Integrated Analysis of the Gut Microbiota and Their Fermentation Products in Mice Treated with the Longevity Enhancing Drug Acarbose

by

Byron J. Smith

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Ecology and Evolutionary Biology)
in the University of Michigan
2019

Doctoral Committee:

Professor Thomas M. Schmidt, Chair
Professor Gregory J. Dick
Professor Meghan A. Duffy
Professor Aaron A. King

Byron J. Smith

bjsm@umich.edu

ORCID iD: 0000-0002-0182-404X

© Byron J. Smith 2019

ACKNOWLEDGMENTS

This work is supported by The Glenn Foundation for Medical Research, the Host Microbiome Initiative at the University of Michigan, and an Integrated Training in Microbial Systems fellowship. Additional funding came from the Department of Ecology and Evolutionary Biology at the University of Michigan.

I want to thank my advisor, Tom, for his wonderful mentorship over the entirety of my PhD. His advising style—usually hands off, and often more of a collaborator than a director of his students' work—has been integral to my graduate school experience. Given the freedom and time to explore the field on my own, I have developed more as an individual scientist than I would have imagined possible. I know that I was not always the easiest trainee to work with, but he has been patient throughout.

Similarly, I need to thank the myriad other mentors, including the members of my committee, whose feedback and optimism have been a vital resource. Aaron, Greg, and Meg have each had important impacts on not only my science, but also how I think about my PhD and my future career. I also want to thank Rich Miller, Pat Schloss, and Nicole Koropatkin for their guidance and feedback at several key points in my research. Likewise, my colleagues in the Schmidt, Schloss, Young, and other MSRB 1 labs, have been friends and collaborators. I would also like to thank everyone who has helped me to navigate the complicated web of paperwork and requirements that is graduate school and the university. I have been fortunate to have support from some of the most talented and helpful administrative staff, including Karrie Black, Michael Dority, Ann Murphy, Becky Mansel, Jane Sullivan, and Cindy Carl.

I would like to in particular acknowledge my wonderful friends in Ann Arbor and East Lansing. You are too numerous to count, but every one of you has contributed to my mental well-being during a long PhD and have made my time in Michigan rewarding personally and not just scientifically. To my cohort and department-mates at (originally) Michigan State University and now the University of Michigan, as well as all of the other ITiMS trainees, your

collegiality and commiseration have been invaluable. In particular, I want to thank Marian Schmidt, my microbial ecologist "sibling" in EEB, who's friendship and feedback over Indian buffets has been a constant. To my roommates at, 2817 Aurora, 529 Elizabeth, and now 1233 Joyce, you have all been the best kinds of friends to come home to; thanks for putting up with both my parties and my idiosyncrasies. To my fēminæ and associated friends, I might hang out with too many women, but I don't regret hanging out with you. Finally, to my frisbee/physics/ski-trip friends—you're hard to contextualize, but you know who you are—I'm lucky to have met all of you, and I look forward to continuing our traditions in the west and far into the future.

I also have to recognize the outsized impact that my family has had on me both before and during graduate school. I would not be the scientist or the human that I am today without them. To my mom, for her support and her influence on my personality. To my dad, for years of curiosity and conversation about the natural world. And to my brother for leading the way in all things academic. I am eternally grateful for your roles in forming me.

Perhaps more than anyone else, I need to thank my partner, Tara, for her unwavering support during my years in Michigan, many of them spent a multi-hour flight away. Her patience and love have been a pillar throughout my PhD. I know that she would have liked for me to finish this thing years ago, but she has been nothing but supportive and proud of me for the duration.

Finally, I would like to thank the murine and bacterial participants in the Interventions Testing Program and other projects at the University of Michigan, without whom my research could never have happened.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

ACA Acarbose

AUC Area under the curve

bp base pair

CBM Carbohydrate binding module

DM Dirichlet-multinomial distribution

FDR False-discovery rate

GH Glycoside hydrolase

GMM Gaussian mixture model

HPLC High-performance liquid chromatography

ITP Interventions Testing Program

LASSO Least absolute shrinkage and selection operation

MAG Metagenome assembled genome

MCC Matthew's correlation coefficient

MCMC Markov chain Monte Carlo

OPF Operational protein family

OTU Operational taxonomic unit

SCFA Short-chain fatty acid

TJL The Jackson Laboratory

UM The University of Michigan

UT The University of Texas Health Science Center at San Antonio

WAIC Widely applicable information criterion

ABSTRACT

During the last two decades, the predominant view of the microbial inhabitants of the mammalian digestive system has evolved from passive commensals to important drivers of health and disease. Processes now known to be affected by the gut microbiome include digestion, immune development and regulation, drug metabolism, pathogen resistance, and many more. Discoveries like these have been driven by revolutionary new methods for the untargeted, high-throughput characterization of the genetic and metabolic composition of microbial communities. However, going from these high-dimensional observations to mechanistic understanding is not trivial and is limited by experimental challenges in studying complex communities in realistic environments. The gut microbiome is particularly difficult, given its taxonomic diversity, physical inaccessibility, and intimate interface with host physiology. In this dissertation, I describe several contributions to our understanding of this important ecological system, with a particular focus on the analysis of bacteria and their metabolic roles *in situ* through the integration of diverse data.

The drug acarbose inhibits the breakdown and absorption of starch in the upper digestive system, resulting in increased availability of this polysaccharide in the lower gut. Interestingly, acarbose has been shown in mice to substantially increase lifespan. This work explores the effects in mice of experimental treatment with acarbose on the composition and function of the gut microbiome. Resulting dramatic increases in the abundance of members of the largely uncultivated bacterial family *Muribaculaceae* are linked to higher concentrations in feces of several short-chain fatty acids—in particular propionate—and these metabolic products of bacterial fermentation are in turn found to be associated with increased mouse lifespan. Furthermore, based

on the culture-free reconstruction of bacterial genomes, we propose a metabolic role of *Muribaculaceae* in the breakdown of starch. Genetic features with homology to the starch utilization system in *Bacteroides* are identified in specific members of this family, possibly explaining their increased abundance in acarbose treated mice. In addition, for one taxon, two distinct genomic variants are found, predicting differences in physiology that could explain variable response to acarbose across replications of the experiment at multiple study sites. Finally, I develop experimental and analysis methods for measurements of absolute abundance in microbial communities using a recently proposed spike-in quantification approach. A novel, model-based inference procedure harnessing these data is found to outperform other methods in identifying changes in bacterial abundance.

This dissertation presents a comprehensive exploration of the dynamics and importance of the gut microbiome in an experimental model with implications for human health. Simultaneously, we develop and refine methods that can be applied to a variety of systems for deriving new understanding about complex microbial communities.

CHAPTER 1

Introduction

The importance of bacterial symbionts in human health and disease is increasingly recognized [1]. Bacteria in the lower gut are now known to play a major role in numerous processes, including digestion [2–4], immune development and regulation [5, 6], drug metabolism [7], pathogen resistance [8], and many more. Given its eminent importance, this microbial system is now referred to as the gut microbiome, and has been described as a "forgotten organ" [9]. These discoveries have in many ways been driven by revolutionary new methods for the untargeted, high-throughput description of the taxonomic, DNA, RNA, protein, and metabolite composition of microbial communities, characterizations which are referred to as metagenomics [10, 11], metatranscriptomics [12], metaproteomics [13], and metabolomics [14].

These "meta-omics" and other modern tools have enabled studies of diverse and ecologically complex bacterial communities. Perhaps the most widely applied of these methods is the 16S rRNA gene survey developed by Norman Pace [15, 16], harnessing the molecular taxonomy made possible by Carl Woese [17]. Using this method, a diversity of bacterial taxa can be simultaneously counted and tracked, despite an inability to cultivate many of these in the lab [18]. The introduction and optimization of this approach has democratized microbial ecology and resulted in an explosion of such studies in previously unconsidered fields, including autoimmune disease [19], obesity [20], autism [21], and longevity [22]. As a result, myriad associations between microbiome composition and various features of host physiology have been described.

However, progress in understanding the mechanistic basis for these associations has not kept pace; demonstrating a causal role of the gut microbiome requires extensive follow-up experimentation. When the relevant bacterial cultivars are not available, one popular approach is the transfer of whole gut communities between animals to identify physiological features that are also transmitted [23]. This method has resulted in perhaps the most important therapy to emerge from the nascent field,

fecal microbiome transplantation, which has demonstrated efficacy as a treatment for *Clostridium difficile* infection [24]. Unfortunately, transplantation and other experimental manipulations are often not feasible, especially in studies of human health, or are not precise enough tools to refine mechanistic understanding. In addition, findings in mouse model systems frequently do not translate to humans [25], in part because of important differences in the microbial communities associated with each host. What's more, microbial community composition can be highly variable between individuals, and given the potential importance of community context on outcomes, this may greatly limit inferences. As a result, deriving actionable insights from studies of the gut microbiome remains a major challenge.

The present dissertation is based on the premise that improved analyses will more efficiently and accurately derive inference from observation, and will facilitate top-down study of the mammalian gut microbiome *in vivo*, in order to better understand its ecology and role in host health. New methods for the analysis of microbial community data have the potential to expand the impact of modern tools by combining multiple data types, like taxonomic surveys, metagenomic sequence, and metabolite concentrations, and can enable new insights into the biology of microbial communities when manipulations are infeasible and model organisms are insufficient. In addition, by complementing improved analyses with experimental perturbations, we may gain greater understanding of the gut microbiome.

Chapter 2 characterizes the impact of the anti-diabetic drug acarbose on the composition of the mouse gut bacterial community and its fermentation products. It explores the possibility that the increased production by gut bacteria of short-chain fatty acids in treated mice results in extended lifespan, potentially explaining the observation of longevity enhancement with the drug [26]. Chapter 3 expands on this with an analysis of 8 reconstructed genomes from members of this bacterial community, all in the largely uncultured family *Muribaculaceae* [27]. By comparing the functional potential of taxa that respond positively to acarbose treatment to those that do not, the ecological niche of this clade is better defined. In Chapter 4 a recently developed extension to the marker gene survey is discussed. Spike-in quantification has the potential to democratize the measurement of absolute abundance in bacterial communities, much as the 16S rRNA gene survey did originally for relative abundance. This chapter presents an overview and suggests several best practices for the approach. In Chapter 5, a statistical model for spike-in quantification data is presented and applied to both simulated and real data. This chapter presents a foundation for future extensions to the model, which can leverage spike-in experiments for improved

ecological inference.

This dissertation presents a comprehensive approach to the analysis of microbial community data that can be applied across systems. The current deluge of metaomics data has the potential to revolutionize the field when combined with carefully designed experimentation and core knowledge about the physiology and ecology of bacteria. The work described here builds on recent advances in microbial ecology, and contributes both new tools and new biological understanding.

BIBLIOGRAPHY

[1] Young, V.B.: The role of the microbiome in human health and disease: An introduction for clinicians. BMJ (Online) **356** (2017). doi:10.1136/bmj.j831

[2] El Kaoutari, A., Armougom, F., Gordon, J.I., Raoult, D., Henrissat, B.: The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. Nature reviews. Microbiology **11**(7), 497–504 (2013). doi:10.1038/nrmicro3050

[3] Cantarel, B.L., Lombard, V., Henrissat, B.: Complex carbohydrate utilization by the healthy human microbiome. PLoS ONE **7**(6), 1–10 (2012). doi:10.1371/journal.pone.0028742

[4] Bäckhed, F., Ley, R.E., Sonnenburg, J.L., Peterson, D.A., Gordon, J.I.: Host-Bacterial mutualism in the human intestine. Science **307**(5717), 1915–1920 (2005). doi:10.1126/science.1104816

[5] Markle, J.G.M., Frank, D.N., Adeli, K., Von Bergen, M., Danska, J.S.: Microbiome manipulation modifies sex-specific risk for autoimmunity. Gut Microbes **5**(4), 485–493 (2014). doi:10.4161/gmic.29795

[6] Brestoff, J.R., Artis, D.: Commensal bacteria at the interface of host metabolism and the immune system. Nature Immunology **14**(7), 676–684 (2013). doi:10.1038/ni.2640

[7] Wilson, I.D., Nicholson, J.K.: Gut microbiome interactions with drug metabolism, efficacy, and toxicity. Translational Research **179**, 204–222 (2017). doi:10.1016/j.trsl.2016.08.002

[8] Britton, R.A., Young, V.B.: Interaction between the intestinal microbiota and host in Clostridium difficile colonization resistance. Trends in Microbiology **20**(7), 313–9 (2012). doi:10.1016/j.tim.2012.04.001

[9] O'Hara, A.M., Shanahan, F.: The gut flora as a forgotten organ. EMBO Reports **7**(7), 688–693 (2006). doi:10.1038/sj.embor.7400731. NIHMS150003

[10] Thomas, T., Gilbert, J., Meyer, F.: Metagenomics - a guide from sampling to data analysis. Microbial informatics and experimentation **2**(1), 3 (2012). doi:10.1186/2042-5783-2-3

[11] Handelsman, J.: Metagenomics: Application of Genomics to Uncultured Microorganisms. Microbiology and Molecular Biology Reviews **68**(4), 669–685 (2004). doi:10.1128/MMBR.68.4.669-685.2004

[12] Bashiardes, S., Zilberman-Schapira, G., Elinav, E.: Use of metatranscriptomics in microbiome research. Bioinformatics and Biology Insights **10**, 19–25 (2016). doi:10.4137/BBI.S34610

[13] Kleiner, M., Thorson, E., Sharp, C.E., Dong, X., Liu, D., Li, C., Strous, M.: Assessing species biomass contributions in microbial communities via metaproteomics. Nature Communications **8**(1) (2017). doi:10.1038/s41467-017-01544-x

[14] Martin, F.P.J., Wang, Y., Sprenger, N., Yap, I.K.S., Rezzi, S., Ramadan, Z., Peré-Trepat, E., Rochat, F., Cherbut, C., Van Bladeren, P., Fay, L.B., Kochhar, S., Lindon, J.C., Holmes, E., Nicholson, J.K.: Top-down systems biology integration of conditional prebiotic modulated transgenomic interactions in a humanized microbiome mouse model. Molecular Systems Biology **4**(205) (2008). doi:10.1016/j.sasoi.2016.05.010

[15] Lane, D.J., Pace, B., Olsen, G.J., Stahl, D.A., Sogin, M.L., Pace, N.R.: Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proceedings of the National Academy of Sciences **82**(20), 6955–6959 (1985). doi:10.1073/pnas.82.20.6955

[16] Schmidt, T.M., DeLong, E.F., Pace, N.R.N., Biology, C.: Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. Journal of bacteriology **173**(14), 4371–4378 (1991). doi:10.1128/jb.173.14.4371-4378.1991

[17] Woese, C.R., Fox, G.E.: Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proceedings of the National Academy of Sciences **74**(11), 5088–5090 (1977). doi:10.1073/pnas.74.11.5088. arXiv:1011.1669v3

[18] Walker, A.W., Duncan, S.H., Louis, P., Flint, H.J.: Phylogeny, culturing, and metagenomics of the human gut microbiota. Trends in Microbiology **22**(5), 267–274 (2014). doi:10.1016/j.tim.2014.03.001. arXiv:1011.1669v3

[19] Markle, J.G.M., Frank, D.N., Mortin-toth, S., Robertson, C.E., Feazel, L.M., Rolle-kampczyk, U., Bergen, M.V., Mccoy, K.D., Macpherson, A.J., Danska, J.S.: Sex differences in the gut microbiome drive hormone-dependent regulation of autoimmunity. Science **339**(March), 1084–1088 (2013). doi:10.1126/science.1233521

[20] Ley, R.E., Turnbaugh, P.J., Klein, S., Gordon, J.I.: Human gut microbes associated with obesity. Nature **444**(21), 1022–1023 (2006). doi:10.1038/4441022a

[21] Sampson, T.R., Mazmanian, S.K.: Control of brain development, function, and behavior by the microbiome. Cell Host and Microbe **17**(5), 565–576 (2015). doi:10.1016/j.chom.2015.04.011. 15334406

[22] Heintz, C., Mair, W.: You are what you host: Microbiome modulation of the aging process. Cell **156**(3), 408–411 (2014). doi:10.1016/j.cell.2014.01.025

[23] Goodrich, J., Waters, J., Poole, A., Sutter, J., Koren, O., Blekhman, R., Beaumont, M., Van Treuren, W., Knight, R., Bell, J., Spector, T., Clark, A., Ley, R.: Human Genetics Shape the Gut Microbiome. Cell **159**(4), 789–799 (2014). doi:10.1016/j.cell.2014.09.053

[24] Rubin, T.a., Gessert, C.E., Aas, J., Bakken, J.S.: Fecal microbiome transplantation for recurrent Clostridium difficile infection: Report on a case series. Anaerobe (November), 1–5 (2012). doi:10.1016/j.anaerobe.2012.11.004

[25] Nguyen, T.L.A., Vieira-Silva, S., Liston, A., Raes, J.: How informative is the mouse for human gut microbiota research? Disease Models & Mechanisms **8**(1), 1–16 (2015). doi:10.1242/dmm.017400

[26] Harrison, D.E., Strong, R., Allison, D.B., Ames, B.N., Astle, C.M., Atamna, H., Fernandez, E., Flurkey, K., Javors, M.A., Nadon, N.L., Nelson, J.F., Pletcher, S., Simpkins, J.W., Smith, D.L., Wilkinson, J.E., Miller, R.A.: Acarbose, 17-α-estradiol, and nordihydroguaiaretic acid extend mouse lifespan preferentially in males. Aging Cell **13**(2), 273–282 (2014). doi:10.1111/acel.12170

[27] Lagkouvardos, I., Pukall, R., Abt, B., Foesel, B.U., Meier-Kolthoff, J.P., Kumar, N., Bresciani, A., Martínez, I., Just, S., Ziegler, C., Brugiroux, S., Garzetti, D., Wenning, M., Bui, T.P.N., Wang, J., Hugenholtz, F., Plugge, C.M., Peterson, D.A., Hornef, M.W., Baines, J.F., Smidt, H., Walter, J., Kristiansen, K., Nielsen, H.B., Haller, D., Overmann, J., Stecher, B., Clavel, T.: The Mouse Intestinal Bacterial Collection (miBC) provides host-specific insight into cultured diversity and functional potential of the gut microbiota. Nature Microbiology **1**(August), 16131 (2016). doi:10.1038/nmicrobiol.2016.131

CHAPTER 2

Changes in the gut microbiota and fermentation products associated with enhanced longevity in acarbose-treated mice.

A version of this chapter has been submitted for publication and has been made available online as [1]

Abstract

Background Treatment with the α-glucosidase inhibitor acarbose increases median lifespan by approximately 20% in male mice and 5% in females. This longevity extension differs from dietary restriction based on a number of features, including the relatively small effects on weight and the sex-specificity of the lifespan effect. By inhibiting host digestion, acarbose increases the flux of starch to the lower digestive system, resulting in changes to the gut microbiota and their fermentation products. Given the documented health benefits of short-chain fatty acids (SCFAs), the dominant products of starch fermentation by gut bacteria, this secondary effect of acarbose could contribute to increased longevity in mice. To explore this hypothesis, we compared the fecal microbiome of mice treated with acarbose to control mice at three independent study sites.

Results Microbial communities and the concentrations of SCFAs in the feces of mice treated with acarbose were notably different from those of control mice. At all three study sites, the bloom of a single bacterial taxon was the most obvious response to acarbose treatment. The blooming populations were classified to the largely uncul-

tured *Bacteroidales* family *Muribaculaceae* and were the same taxonomic unit at two of the three sites. Total SCFA concentrations in feces were increased in treated mice, with increased butyrate and propionate in particular. Across all samples, *Muribaculaceae* abundance was strongly correlated with propionate and community composition was an important predictor of SCFA concentrations. Cox proportional hazards regression showed that the fecal concentrations of acetate, butyrate, and propionate were, together, predictive of mouse longevity even while controlling for sex, site, and acarbose.

Conclusion We have demonstrated a correlation between fecal SCFAs and lifespan in mice, suggesting a role of the gut microbiota in the longevity-enhancing properties of acarbose. Treatment modulated the taxonomic composition and fermentation products of the gut microbiome, while the site-dependence of the microbiota illustrates the challenges facing reproducibility and interpretation in microbiome studies. These results motivate future studies exploring manipulation of the gut microbial community and its fermentation products for increased longevity, and to test a causal role of SCFAs in the observed effects of acarbose.

2.1 Background

The Interventions Testing Program (ITP) is a long-running, well-powered study of longevity enhancing interventions in genetically heterogeneous mice with identical protocols replicated at each of three study sites [2]. The drug acarbose (ACA) has been reproducibly shown in that study to increase mouse median lifespan with a larger effect in males than females [3, 4]. The largest increase was found when treatment began at 4 months, 22% in males and 5% in females [3], but the beneficial effect was still detectable in mice receiving ACA starting at 16 months [4]. The 90th percentile lifespan, a surrogate for maximum lifespan also shows benefits of ACA, with similar magnitudes in both male and female mice [3]. ACA is a competitive inhibitor of α-glucosidase and α-amylase, resulting in delayed intestinal breakdown of starch when taken with food and reduced postprandial increases in blood glucose. For these reasons, ACA is prescribed for the treatment of type 2 diabetes mellitus [5], and has also been shown to reduce the risk of cardiovascular disease in that population [6].

It is unclear whether the pathways by which ACA extends longevity in mice overlap with those affected by dietary restriction, but several observations have suggested critical differences [3]. Weight loss in ACA mice was more dramatic in females than

in males, while the longevity effect is much stronger in males. By contrast, dietary restriction affects both weight and lifespan similarly in both sexes. Likewise, the response of fasting hormone FGF21 to ACA treatment was opposite in direction from that induced by dietary restriction. In female mice alone, ACA blocked age-related changes in spontaneous physical activity, while dietary restriction leads to dramatic increases in activity in both sexes. It is therefore justified to suspect that the effects of ACA on longevity are due to pathways distinct from dietary restriction.

Besides reducing the absorption of glucose from starch, inhibition of host enzymes by ACA results in increased flow of polysaccharide substrate to the lower digestive system [7]. ACA has been shown to raise the concentration of starch in stool [7, 8], and the observed increased excretion of hydrogen in breath [9–13] demonstrates that at least some of this substrate is fermented by the gut microbiota. The major byproducts of polysaccharide fermentation in the gut are hydrogen, CO_2, and short-chain fatty acids (SCFAs), in particular acetate, butyrate, and propionate. Unsurprisingly, ACA treatment has been observed to increase acetate concentrations in human feces [14] and serum [15], as well as concentrations in portal blood and total amounts in rodent cecal contents [7]. Likewise, in some studies, ACA increased butyrate concentrations in human feces [8, 13, 16] and serum [15]. ACA also increased propionate concentrations in rat portal blood and total amount in cecal contents [7], as well as total output in feces in humans [14]. These changes were presumably due to changes in the activity and composition of microbial fermenters in the lower gut. Indeed, ACA was found to modulate the composition of the fecal bacterial community in prediabetic humans [17], and both increase the SCFA production potential inferred from metagenomes and lower fecal pH [18].

Impacts of ACA on microbial fermentation products are of particular interest because SCFAs produced in the gut are known to affect host physiology, with a variety of health effects associated with butyrate, propionate, and acetate (reviewed in [19] and [20]). Although butyrate and propionate are primarily consumed by the gut epithelium and liver, respectively, they are nonetheless detectable in peripheral blood, and acetate can circulate at substantially higher concentrations [21]. Four G-protein coupled receptors have been shown to respond to SCFAs with varying levels of specificity: FFAR2, FFAR3, HCA2, and OLFR78. All except OLFR78 are expressed in colonic epithelial cells, and each is expressed in a variety of other tissues throughout the body. Similarly, both butyrate and propionate act as histone deacetylase inhibitors which could have broad effects on gene expression through modulation of chromatin structure. In total, these pathways contribute to regulation of cellular

proliferation, inflammation, and energy homeostasis, among other processes. The effects of ACA on fermentation products in the gut may, therefore, modulate its overall effects on host physiology.

Despite theoretical expectations and suggestive empirical results in other animal models, no study has looked for direct evidence that some of the longevity enhancing effects of ACA in mice are mediated by the gut microbiota and the SCFAs produced during fermentation. Here we test four predictions of that hypothesis: (1) ACA reproducibly modulates bacterial community composition; (2) the concentrations of SCFAs are increased in ACA-treated mice; (3) community structure is correlated with SCFAs and other metabolites in both control and treated mice; and (4) SCFA concentrations are predictive of lifespan. Fecal samples were analyzed from control and ACA treated mice enrolled in the ITP protocol at three, independent study sites.

2.2 Results

2.2.1 Study population

Sampled mice are representative of an underlying study population that recapitulates the previously observed sex-specific longevity effects of ACA. Across all three sites, ACA increased the median male survival of the underlying study population by 17% from 830 to 975 days (log-rank test $P < 0.001$). Female median survival increased 5% from 889 to 931 days ($P = 0.003$). These results are consistent with the increased longevity due to ACA previously reported [3]. Fecal samples from 48 mice at each of three study sites—The Jackson Laboratory (TJL), the University of Michigan (UM), and the University of Texas Health Science Center at San Antonio (UT)—were collected between 762 and 973 days of age with a balanced factorial design over sex and treatment group. Visual inspection of the overall survival curves (Figure 2.1) confirmed that the longevity of mice sampled for microbiome analyses at UM and UT was representative of the other, surviving, unsampled mice. Samples from TJL were not matched to individual mice and longevity measures are not available for the subset described here.

2.2.2 Differences in fecal community in ACA-treated mice

ACA-treated mice had a substantially different microbial community composition from control mice at all three study sites. In a multivariate analysis of variance on site, sex, and treatment using Bray-Curtis dissimilarities and including all two-way

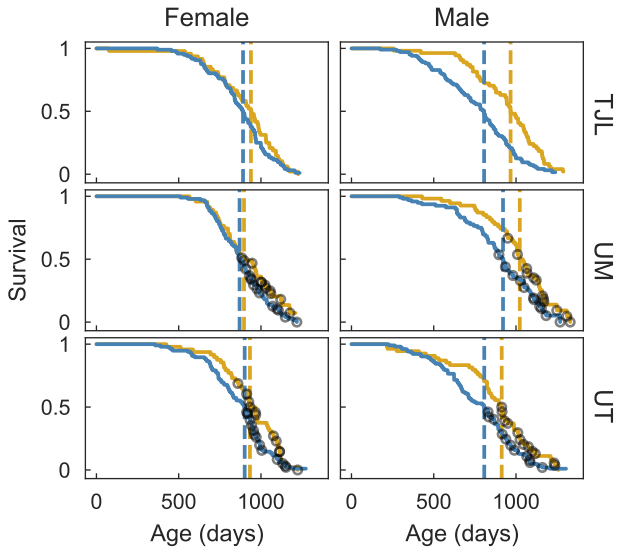

Figure 2.1: Survival curves for mice treated with acarbose or controls. Fraction of mice surviving on the control diet (blue lines) or mice fed the same diet containing ACA (gold) at each of three sites: TJL, UM, and UT. Median longevity for each group of mice is indicated by a dashed vertical line. Black circles indicate the age at death for each of the sampled mice at UM and UT.

interactions, significant effects were found for treatment (partial $r^2 = 9.6\%$, PERMANOVA $P < 0.001$) and site (partial $r^2 = 16.4\%$, $P < 0.001$), as well as their interaction (partial $r^2 = 3.4\%$, $P < 0.001$). These statistical results reflect the separation apparent in a principal coordinates ordination (see Figure 2.2). A much smaller but still significant effect of sex (partial $r^2 = 1.0\%$, $P = 0.014$) was also identified, but there was no interaction between sex and treatment ($P = 0.344$). Despite the unbalanced design, significance of the PERMANOVA was not affected by changing the order of predictors. Based on a test of multivariate homogeneity of variances, dispersion differed between sites (PERMDISP $P < 0.001$) and sexes ($P = 0.023$), which may bias the PERMANOVA results, but did not differ between treatments ($P = 0.425$). The small effect of sex and the lack of significant interaction effects with treatment suggest that community composition itself, at the level of overall diversity, is not directly responsible for differential effects of ACA on longevity in male and female mice. However, the substantial differences in community composition due to treatment, while not surprising, suggests that the effects of ACA on the microbiome have the potential to modulate host health.

The fecal bacterial community in control mice was dominated by a handful of bacterial families (see Table 2.1 for details). Across control mice at all three sites,

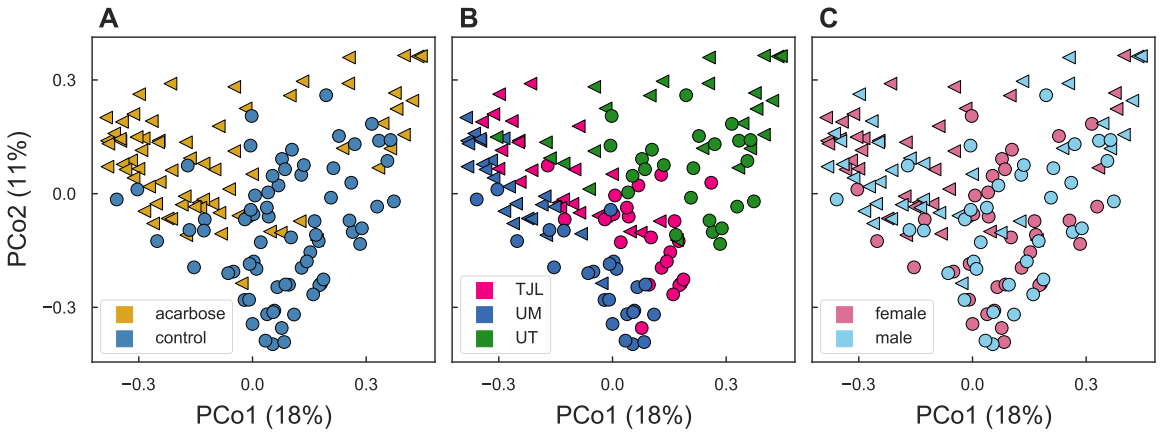

Figure 2.2: Fecal bacterial community composition in sampled mice. The two dominant principal coordinates, based on Bray-Curtis dissimilarities among community profiles, are plotted, and percent of variation explained by each is indicated in parentheses on the axes. The location of points in each panel is identical. Markers denote whether mice were treated (triangles) or controls (circles). In (**A**) points are colored by treatment: control mice (blue) and ACA-treated (gold), in (**B**) points are colored by site: TJL (pink), UM (blue), and UT (green), and in (**C**) points are colored by sex: male (light blue) and female (pink).

a median of 30% of sequences were classified as members of the largely uncultured family *Muribaculaceae*—historically called the S247—belonging to the phylum *Bacteroidetes*. Other abundant families included the *Lachnospiraceae* (27%), *Ruminococcaceae* (14%), *Lactobacillaceae* (9%), and *Erysipelotrichaceae* (1%), all of which are classified in the phylum *Firmicutes*. More than 99.99% of sequences across all mice were classified at or below the family level.

At a 97% sequence similarity cutoff, 271 operational taxonomic units (OTUs) had a mean relative abundance across all samples of greater than 0.01% and an incidence of greater than 5%. Of these, the relative abundance of 113 OTUs differed between treated and control mice, correcting for a false discovery rate (FDR) of 5%. Together, these OTUs account for a median relative abundance of 54% across both control and treated mice. OTUs differing between sexes or reflecting an interaction between sex and treatment were a substantially less abundant. 5 OTUs were identified after FDR correction that differed significantly in relative abundance between male and female mice, accounting for a median, summed relative abundance of 6%. 7 OTUs were found to be subject to an interaction between treatment and sex, with a median relative abundances of 2%.

Differences between control and ACA mice at TJL and UM were dominated by the

increased abundance of a single OTU, OTU-1, which had a median relative abundance of 7.7% in control mice compared to 28.8% in ACA mice at TJL (Mann-Whitney U test $P < 0.001$), and 10.4% compared to 39.0% at UM ($P < 0.001$) (see Figure 2.3). At UT, OTU-1 was higher in ACA-treated mice—a median of 5.4% and 11.0% in control and treated mice, respectively—but this increase was not statistically significant ($P = 0.344$). Instead, a different OTU, designated OTU-4, was strongly affected by ACA treatment at UT, with a median relative abundance of 6.3% in control mice that increased to 25.6% in ACA-treated mice ($P = 0.007$). OTU-4 was nearly absent at TJL and UM, with only one mouse out of 95 having a relative abundance above 0.1%, compared to 39 out of 48 mice at UT. Differences in abundance between sexes were not observed for OTU-1 at TJL or UM, but at UT results were suggestive of an increased abundance of OTU-1 in females ($P = 0.076$) and an increased abundance of OTU-4 in males ($P = 0.060$). Interestingly, their combined abundance did not differ between males and females ($P = 0.344$) at UT. Both OTU-1 and OTU-4 were classified as members of the *Muribaculaceae*, and subsequent phylogenetic analysis confirmed this placement (Appendix A). OTU-1 and OTU-4 are approximately 90% identical to each other and to the most closely related cultivar (DSM-28989) over the V4 hypervariable region of the 16S rRNA gene. These OTUs are notable both for their high abundance overall, as well as the large difference between control and ACA-treated mice. It is surprising that OTU-4 is common and differentially abundant at UT, while remaining rare at both of the other sites, suggesting that local community composition modulates the effects of ACA. While OTU-1 is made up of multiple unique sequences, the composition within the OTU does not differ substantially with ACA treatment (see Figure 2.4).

The increased relative abundance of OTU-1 and OTU-4 in mice treated with ACA appears to be due to greater abundance of these sequences, and is not explained solely by a decrease in other groups. The abundance of taxa in control and treated mice was compared based on the recovery of spiked-in standard relative to the sequence of interest. The median combined spike-adjusted abundance of 16S rRNA gene copies from OTU-1 and OTU-4 was 4.3 times greater per gram of feces in ACA-treated mice compared to controls (Mann-Whitney U test $P < 0.001$), suggesting a corresponding increase in population density.

While OTU-1 and OTU-4 are classified to the same family and are similarly affected by ACA, other OTUs in the *Muribaculaceae* have decreased abundance in treated mice. The combined relative abundance of all other OTUs in the family—excluding OTU-1 and OTU-4—was 8.3% in treated mice versus 16.8% of sequences

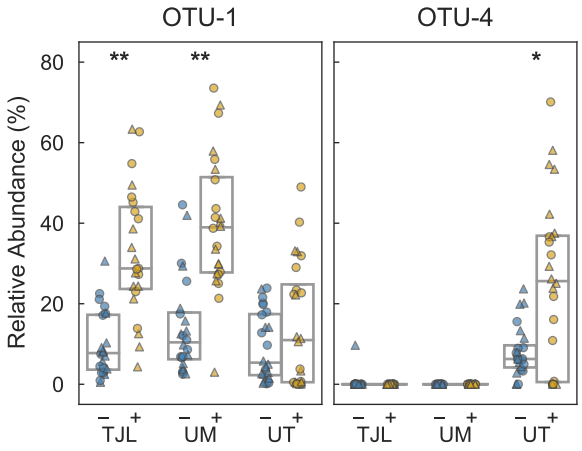

Figure 2.3: Abundance of two dominant OTUs in feces of sampled mice. Relative abundance of the 16S rRNA gene from OTU-1 and OTU-4 in ACA-treated mice (gold) compared to controls (blue). Points in each panel correspond with samples collected from individual mice at each of three replicate study sites. Markers indicate the sex of the mouse: male (triangle) or female (circle). Boxes span the interquartile range and the internal line indicates the median. (*: $P < 0.05$, **: $P < 0.001$ by Mann-Whitney U test).

in control mice (Mann-Whitney U test $P < 0.001$). The median combined spike-adjusted abundance of all other *Muribaculaceae* OTUs was 0.5 times the median in control mice ($P = 0.001$), suggesting a decrease in the population density of these taxa. This is consistent with competition between OTUs in this family.

Three of the five most abundant families all exhibit decreased relative abundance in ACA treated mice (see Table 2.1). However, the large increase in abundance of OTU-1 and OTU-4 suggests that some changes in the relative abundance of other taxa may be the result of compositional effects, rather than decreased density. For instance, although the relative abundance of *Ruminococcaceae* was lower in ACA-treated mice, the spike-adjusted abundance was little changed ($P = 0.327$), emphasizing the value of this complementary analysis. Conversely, decreased relative abundance was matched by decreased spike-adjusted abundance for both the *Lactobacillaceae* ($P = 0.014$) and the *Erysipelotrichaceae* ($P = 0.063$).

ACA-treated mice exhibited decreased fecal community diversity. The median Chao1 richness estimate was decreased from 229 in control mice to 199 in treated mice (Mann-Whitney U test $P < 0.001$). The Simpson's evenness index was also lower in ACA mice: 0.044 versus 0.075 in controls ($P < 0.001$). This reduced richness and evenness is not surprising given the much greater abundance of a single OTU in treated mice at each site. To understand changes in diversity while controlling for

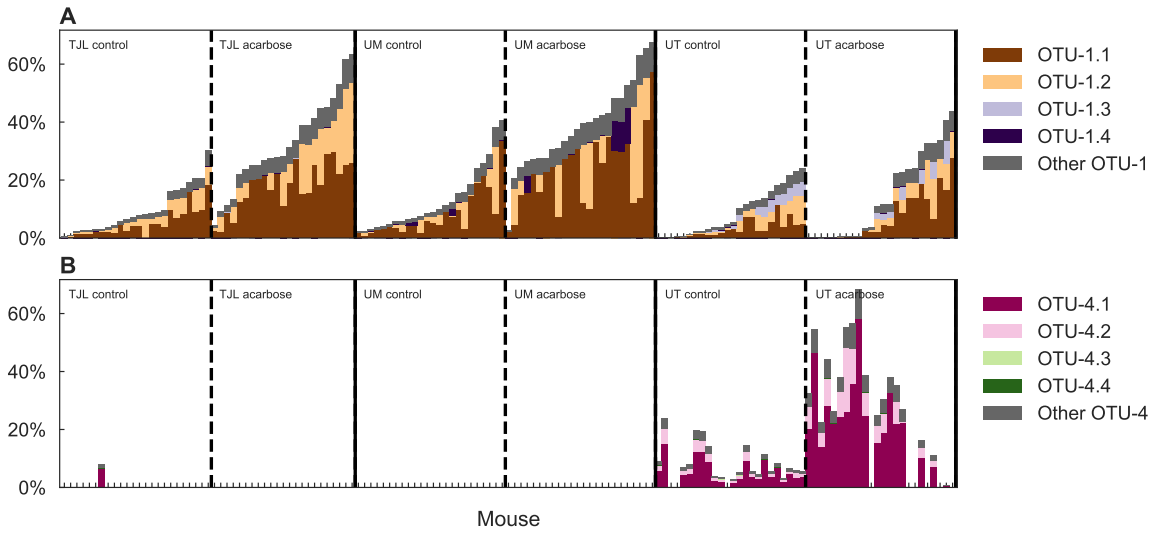

Figure 2.4: Relative abundance of unique 16S rRNA gene sequences in two OTUs. Stacked bars indicate the composition of sequences clustered into (**A**) OTU-1 and (**B**) OTU-4. Colors are assigned to the top four most common sequences within each OTU and all remaining sequences from that OTU are assigned the color gray. Stacked bars in each position represent individual mice sampled for this study, and reflect the relative abundance of unique sequences in that sample. Mice are grouped into sites and then treatments, and finally ordered based on the total abundance of OTU-1.

compositional effects, we measured the effect of ACA ignoring counts for OTU-1 and OTU-4. While Simpson's evenness was not decreased by treatment in this fraction of the community ($P = 0.26$), the Chao1 richness—subsampling to equal counts *after* partitioning—was ($P = 0.005$), suggesting that the bloom of OTU-1 and OTU-4 may have, in fact, resulted in the local extinction of rare community members.

2.2.3 Changes in fecal metabolite concentrations

Long-term ACA treatment affects metabolite profiles, increasing concentrations of the SCFAs in feces (see Figure 2.5). Butyrate concentrations were increased from a median of 3.0 mmols/kg wet weight in control mice to 4.9 in ACA-treated mice (Mann-Whitney U test $P < 0.001$). Propionate concentrations were also increased: a median of 1.1 in controls compared to 2.3 with ACA ($P < 0.001$). Median acetate concentrations were higher, 16.2 mmols/kg versus 12.9 in controls, but a Mann-Whitney U test did not surpass the traditional P-value threshold ($P = 0.073$). The summed concentrations of acetate, butyrate, and propionate was greater in ACA-treated mice, with a median concentration of 25.4 mmols/kg versus 19.0 mmols/kg

Table 2.1: Abundance of common bacterial families in treated and control mice

family	% control[a]	% ACA[a]	ACA : control[b]
Muribaculaceae	30.4 (21.5, 43.3)	48.1** (35.3, 61.8)	1.8**
Lachnospiraceae	26.6 (16.3, 41.6)	23.9† (9.6, 37.4)	1.3
Ruminococcaceae	14.2 (9.0, 19.0)	11.6* (6.9, 15.6)	1.1
Lactobacillaceae	9.5 (1.2, 17.0)	2.6* (1.0, 8.2)	0.31*
Erysipelotrichaceae	1.4 (0.3, 6.2)	0.5* (0.2, 2.2)	0.42†

[a] Median and interquartile range of the relative abundance of the top five most abundant bacterial families in control and ACA-treated mice

[b] the ratio of median spike-adjusted abundances in ACA-treated mice versus control mice

† $P < 0.1$ via Mann Whitney U test

* $P < 0.05$

** $P < 0.001$

in control mice ($P = 0.003$). This confirms our predictions given the expectation of greater availability of polysaccharide substrate for fermentation. Indeed, median glucose concentration was also increased from 5.3 to 10.3 ($P < 0.001$). Concentrations of formate, valerate, isobutyrate, and isovalerate were generally below the detection limit. Fresh pellet weight was increased from a median of 36 to 74 mg ($P < 0.001$) Fecal starch content was not measured, but pellets from ACA-treated mice had a noticeably chalky appearance.

Butyrate as a molar percentage of total SCFA was modestly greater in the ACA mice, a median of 19% in control mice was increased to 22% in treated mice (Mann-Whitney U test $P < 0.001$), as was propionate: 7% in control, 10% in treated ($P = 0.006$), while acetate was decreased from 73% in controls to 66% in treated mice ($P < 0.001$).

In contrast to the three measured SCFAs and glucose, both succinate and lactate concentrations were decreased. Median lactate was decreased from 3.2 mmols/kg in control mice to 1.3 in ACA-treated mice ($P = 0.003$), and succinate from 3.0 to 1.6 ($P < 0.001$). It is surprising that these fermentation intermediates are reduced, given the expected increase in available polysaccharide. It is possible that their concentrations reflect greater consumption in downstream pathways, or perhaps ACA directly inhibits the metabolism and growth of relevant bacteria; such effects have been previously reported for *in vitro* fermentations of starch with human fecal slurries [8].

Differences in SCFAs between sexes are particularly interesting given the greater

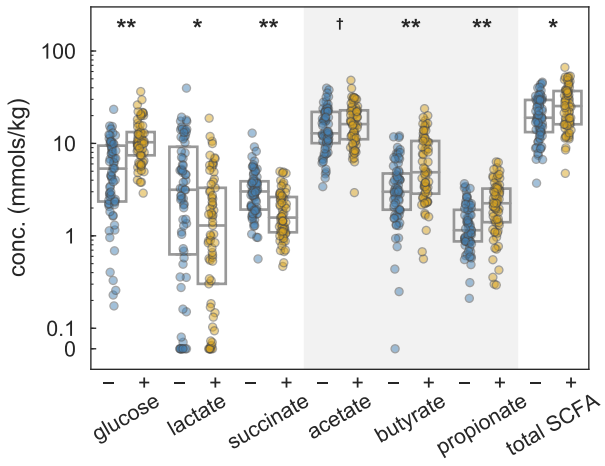

Figure 2.5: Concentrations of metabolites in feces of sampled mice. Feces were obtained from mice fed either the control diet (blue) or the same diet supplemented with ACA (gold). Boxes span the interquartile range and the internal line indicates the median. The shaded region highlights the three major SCFAs produced by microbial fermentation of polysaccharides in the gut, and the sum of their concentrations is plotted as "total SCFA". Above 0.1 mmols/g, concentrations are plotted logarithmically. (†: $P < 0.1$, *: $P < 0.05$, **: $P < 0.001$ by Mann-Whitney U test).

longevity effects of ACA in male mice. For propionate, a sex-by-treatment interaction was found (ANOVA $P = 0.023$), but butyrate and acetate had no such effect. This interaction results in a larger difference in propionate concentrations for male mice (from 1.4 mmols/kg in control to 2.7 in ACA) than for female mice (from 1.0 to 1.9) with ACA treatment. The significance of the interaction term was not corrected for multiple testing, and therefore additional studies would greatly increase our confidence in this result.

2.2.4 Community predictors of fecal SCFA concentrations

Community composition was correlated with metabolite concentrations in both control and ACA-treated mice. Numerous strong correlations were detected between the spike-adjusted abundance of 16S rRNA copies from the most common bacterial families and the concentrations of SCFAs and lactate. Notably, *Muribaculaceae* abundance was particularly strongly correlated with propionate concentrations in both control (Spearman's $\rho = 0.36$, $P = 0.002$; see Figure 2.6) and ACA mice ($\rho = 0.64$, $P < 0.001$). Likewise, *Lachnospiraceae* were correlated with butyrate ($\rho = 0.61$ in control and 0.77 in ACA, $P < 0.001$ for both), and *Lactobacillaceae* with lactate concentrations ($\rho = 0.63$ in control and 0.67 in ACA, $P < 0.001$ for both). Strik-

ingly, concentrations of acetate and butyrate were especially correlated with each other ($\rho = 0.67$ in control and 0.80 in ACA, $P < 0.001$ for both). Although our study was not an unambiguous test, these results support the hypothesis that the fecal metabolite response to treatment is dependent on the population density of relevant microbes in the gut community. Similarly, environmental and host factors that promote or inhibit the growth of particular community members would be expected to modulate the response.

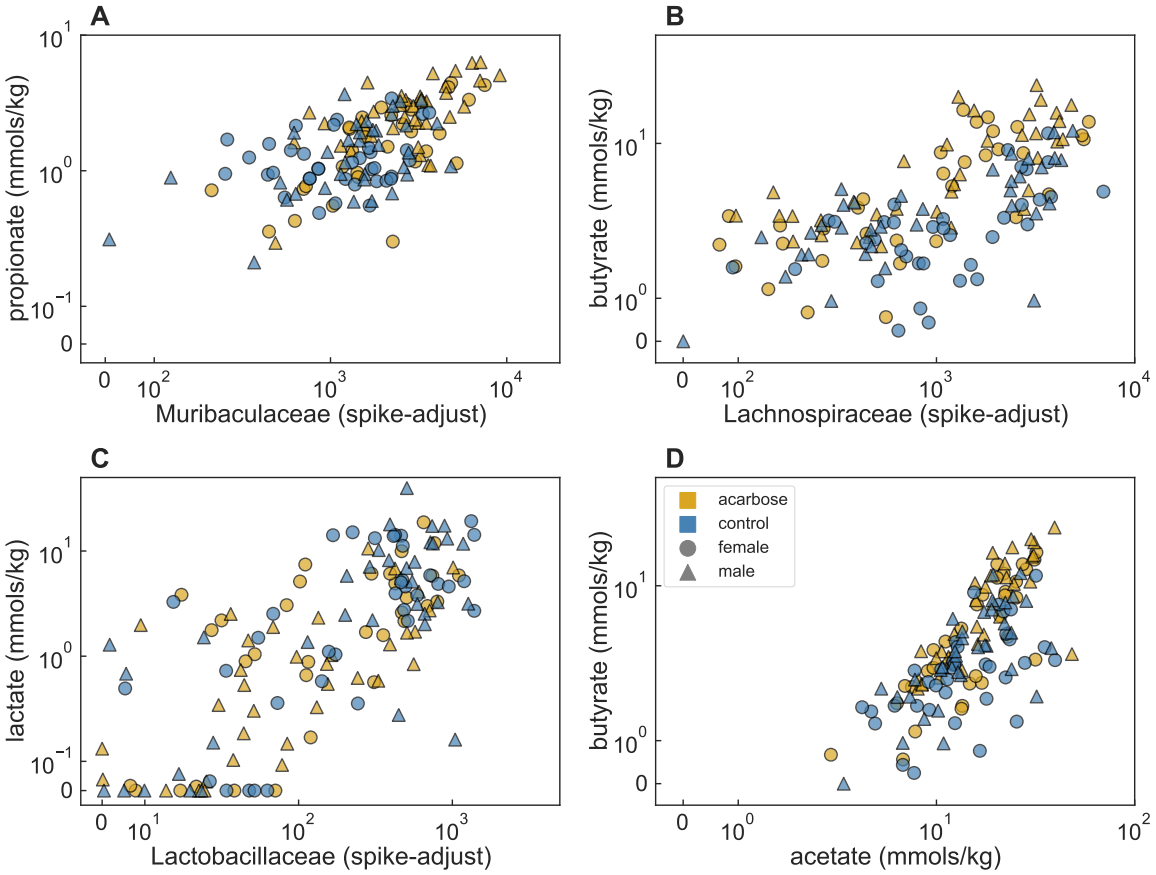

Figure 2.6: Correlations between abundances of taxa and metabolites in feces. Correlations are illustrated among metabolite concentrations in feces and family level, summed spike-adjusted 16S rRNA gene abundances. Points correspond with samples collected from individual mice and colors indicate whether they were obtained from mice fed the control diet (blue) or the same diet supplemented with ACA (gold). Markers indicate the sex of the mouse: male (triangle) or female (circle). Metabolite concentrations are reported normalized to feces wet weight, and abundances are in spike-equivalent units. Values are on a linear scale between 0 and the subsequent tick label, above which, points are plotted logarithmically.

To identify key players in these associations, we examined the relationship be-

tween metabolite concentrations and the spike-adjusted abundances of OTUs. Based on a LASSO regression, the abundances of a number of OTUs can be used to predict concentrations of propionate, butyrate, acetate, and lactate even after accounting for treatment, sex, and study site. Consistent with the correlations found between *Muribaculaceae* abundance and propionate, OTU-1 and OTU-4 were identified as predictors of increased propionate, along with a third taxon, OTU-5, also classified as a member of the family. For both butyrate and acetate, OTUs classified as members of both the *Lachnospiraceae* and *Ruminococcaceae* were most predictive of increased concentrations. Unsurprisingly, the most abundant OTU classified as a member of the *Lactobacillaceae*, OTU-2, was found to be highly predictive of increased lactate. However, 8 OTUs were also associated with decreased lactate concentrations, most of which were among those associated with increased butyrate and acetate. Among other explanations, this is consistent with these taxa either being inhibited by lactate or being lactate utilizers, which are likely to be producing SCFAs as secondary fermentation products. Overall, results were both consistent with *a priori* expectations, and useful for generating hypotheses about which taxa might be associated with the generation of fermentation products.

2.2.5 Fecal SCFA concentrations as predictors of longevity

Given the documented health benefits of SCFAs in the gut (reviewed in [19]) and their increased levels in ACA-treated mice, we tested the relationship between the acetate, butyrate, and propionate concentrations in feces, and the lifespan of individual mice. Lifespans of fecal donors were not available for mice at TJL, so survival analyses were carried out only with UM and UT mice, and effect sizes are reported for SCFAs as standardized hazard ratios (HRs). Due to the reduced number of mice sampled for this study, data were pooled across sexes and sites. The shared effects of the design parameters—treatment, sex, and study site—on both SCFAs and longevity, were accounted for by including terms for these covariates as well as their two and three-way interactions. Analyses reinforcing our interpretations are discussed in Appendix B. Tested individually against this null model, an association between longevity and propionate was found (standardized HR of 0.727, $P = 0.031$), but no relationship was found with butyrate ($P = 0.240$) or acetate ($P = 0.742$). However, when the model was fit with all three SCFAs simultaneously, each was found to be associated with longevity ($P = 0.012$, 0.030, 0.042 for propionate, butyrate, and acetate, respectively). Coefficients for SCFA covariates in this full model suggest a positive as-

sociation with longevity for both propionate and butyrate (standardized HR of 0.674 and 0.586, respectively). Interestingly, a negative association was found with acetate (standardized HR of 1.576) using this model. The discrepancy between this result and the lack of association when butyrate and acetate are each tested alone likely reflects the strong positive correlation between acetate and butyrate concentrations, masking their individual, opposing associations with longevity. The overall fit of the full model was improved compared to the null model with only design covariates (likelihood ratio test, $P = 0.023$).

2.3 Discussion

ACA, by inhibiting the enzymes responsible for starch degradation in the small intestine, is expected to increase the availability of this polysaccharide to the microbiome. The resulting increase in SCFA production may contribute to the effects of ACA on health. Despite previous observations in humans and rats that ACA results in substantial changes to the community structure [8, 17] and fermentation products [7, 8, 13, 16] of the gut microbiota, a link between these effects on the microbiome and longevity has not been established. Here we present the first study to combine bacterial community surveys with measurement of fecal metabolites in ACA treated mice, as well as the first to pair these data with lifespan, allowing us to explore the role of the microbiome in increased longevity.

Our results confirm all four predictions that we set out to test: ACA was found to affect both (1) the composition of bacterial communities and (2) SCFAs in mouse feces, (3) the abundances of individual taxonomic groups were associated with concentrations of fermentation products, and (4) the concentrations of fecal SCFAs were associated with variation in mouse longevity.

While it is unsurprising that an increased flux of starch to the large intestine affected the gut microbiota and their fermentation products, some changes were especially pronounced. The increased relative abundance in ACA-treated mice of the dominant OTU—OTU-1 at UM and TJL and OTU-4 at UT—was dramatic: one or the other was increased approximately 4-fold at all three sites and in multiple samples more than half of sequences belonged to these OTUs. A cursory BLAST search reveals that sequences identical to OTU-1 have been previously recovered in published studies ([e.g. 22, 23]); in [24] the sequence was found at high relative abundance in the brains of mice that had undergone sepsis. It was notable that OTU-1 did not respond to ACA at UT, while its increased abundance was so striking at UM and TJL. Our

results appear to constrain the potential explanations for this observation. OTU-1 was present and abundant at all three sites; the abundance in control mice was lowest at UT, although the median there was still greater than 5%. While it is not possible with the data presented here to rule out genomic differences of OTU-1 among sites, a similar composition of unique 16S rRNA gene sequences made up this cluster at all three. On the other hand, OTU-4 was at very low abundance, with no reads in a majority of samples, at UM and TJL where OTU-1 did respond to ACA. These results suggest that both OTUs respond to ACA in the same way, with OTU-4, when it is sufficiently abundant, inhibiting the response of OTU-1, potentially through resource competition. Both OTUs are in the same family, the *Muribaculaceae*, but are not the same species or genus by the traditional similarity thresholds, sharing only 90% identity over the sequenced fragment. The differential response of these OTUs among sites illustrates the importance of each site's local "metacommunity" in determining the microbial community's response to environmental perturbations.

Pronounced differences in the resident microbial communities of different hosts may contribute to challenges in translating results from mice and other model organisms to humans. A comparison of bacterial community composition in feces in prediabetic people before and during a 4-week ACA treatment period did not reveal changes of the magnitude reported here [17], although this may reflect the limited duration of treatment. Interestingly, in that study *Lactobacillaceae* abundance increased with ACA, while we observed this family to be depleted in treated mice. The abundance of the *Muribaculaceae* was not reported. Although members of this family are common in mice and have been previously shown to respond to diet, the clade is substantially less abundant in most human samples [25]. However, the prevalence of *Muribaculaceae* may be under-reported in the literature, as the Ribosomal Database Project [26] does not include the family and classification using this database assigns sequences to the *Porphyromonadaceae* instead [27]. Historically, two other names have also been used for this clade: the "S247" (from an early environmental clone [28]), and "*Candidatus Homeothermaceae*" (proposed in [25]). While the isolation of one *Muribaculaceae* cultivar has recently been published, *Muribaculum intestinale* YL27 [29], and as of this writing several draft genomes are available for unpublished isolates (e.g. see whole genome shotgun sequencing projects NWBJ00000000.1 and NZ_NFIX00000000.1), additional cultivars will be vital for understanding the function and ecology of the family. Nonetheless, genomes assembled from metagenomes suggest that populations of *Muribaculaceae* are equipped with fermentation pathways to produce succinate, acetate, and propionate, and that the family is composed of metabolic guilds, each

specializing on the degradation of particular types of polysaccharides: plant glycans, host glycans, and α-glucans [25]. This suggests that the *Muribaculaceae* may occupy a similar set of niches in mice as do *Bacteroides* species in humans. The *Muribaculaceae* and *Bacteroides* are both in the order *Bacteroidales*. *Bacteroides* also specialize in the fermentation of polysaccharides [30], and at least some of the most common species in the human gut are known to produce succinate, acetate, and propionate from the fermentation of polysaccharides [19, 31–33]. Unlike the patterns observed here for *Muribaculaceae* in mice, the abundance of *Bacteroides* decreased with ACA treatment in one study in humans [17], suggesting that the microbially mediated effects of ACA may fundamentally differ between these hosts.

Besides hypotheses based on genome content, the correlation between total *Muribaculaceae* abundance and propionate concentrations and the specific association with OTU-1 and OTU-4 found in the LASSO analysis suggest that both OTUs, and perhaps other *Muribaculaceae* species in this study, ferment starch to propionate. This also supports the hypothesis, discussed above, that both OTUs occupy overlapping niches. Although increased butyrate concentrations have been frequently reported with ACA treatment [7, 8, 13, 15, 16], elevated concentrations of propionate have been observed in just one previous study using portal blood in rats [7]. Studies in humans have instead found decreased or no change in fecal [8, 13, 14] or serum [15] propionate concentrations with ACA. Decreased propionate has been attributed to preferential production of butyrate from starch fermentation [34, 35] or inhibition of propiogenic bacteria by ACA [8]. Our observation of increased propionate was robust and reproduced at all three sites. If this conflicting result reflects both the greater initial abundance and enrichment in our study of the *Muribaculaceae*—especially OTU-1 and OTU-4—it demonstrates the value of measuring both community composition and metabolite concentrations in the same samples.

SCFAs are commonly suggested to act as intermediaries between the gut microbiota and host physiology [19]. While our study was not designed to provide a causal test of effects of SCFAs on longevity, and the power of our analysis was limited, a statistical association between SCFA concentrations and mouse lifespan supports an interpretation that is consistent with an extensive literature on the health benefits of butyrate and propionate [19]. In addition, that SCFA concentrations were associated with longevity above and beyond the effects of ACA, study site, and sex, further supports this hypothesis. It is somewhat surprising, however, that a single fecal sample taken, in some cases, several months before death, could be predictive of longevity. The association reported here could reflect other, unmeasured, changes in the gut

microbiome or host physiology. Concentrations of metabolites in feces are an integration of both production and consumption rates along the length of the lower gut, and may not reflect host exposure nor the strength of host physiological response. It is also important to note that, since all mice in this study at the time of sampling were of an age close to the median lifespan of control individuals, the results are only relevant to mechanisms of aging in late-life and should not be extrapolated to young mice. Experimental tests of a causal role for SCFAs in longevity will be challenging, as they likely require controlled manipulation of intestinal SCFAs for the lifetime of a mouse.

Due to the preferential enhancement of longevity by ACA in male mice we sought to identify aspects of the gut microbiome that responded differently in male and female mice (i.e. interaction effects), as these might suggest mechanistic explanations for differences in longevity effects [3]. While we do not believe that the magnitude of sex-by-treatment interactions observed for various aspects of the microbiome were sufficiently pronounced to fully explain the differential effect of ACA on lifespan, our search was limited by sample size, variability between study sites, and the large number of features being tested. Nonetheless, ACA was found to increase propionate concentrations more in male mice than females, a statistically significant pattern before correction for multiple testing, and the relative abundance of a handful of OTUs seem to have also been subject to an interaction.

Mechanisms unrelated to the gut microbiome have been proposed for the effects of ACA on lifespan. Because ACA reduces the postprandial glucose spike observed in mice and humans, hypotheses emphasizing the reduction of harmful effects associated with these transient surges have been most commonly invoked. Studies of UM-HET3 mice given ACA from 4 to 9 months of age suggested that mean daily blood glucose levels are minimally affected, but that absorption of glucose was both slower and longer lasting [3]. Interestingly, fasting insulin level in ACA-treated males are much lower than those in control males, consistent with an improvement in insulin sensitivity [3]. This reduction was not seen in females, where insulin levels in controls were lower than in control males and similar to those in ACA-treated males [3, 36], presenting one possible explanation for the stronger longevity benefit in males. Still, the connection between this modulation of postprandial glucose—with or without improved insulin sensitivity—and extended longevity is still far from certain.

The work presented here explores a different hypothesis: that health benefits in ACA mice are related to changes in the activity of microbial communities in the gut associated with the increased influx of starch, and possibly attributable to known

health effects of microbial metabolites, including SCFAs. The changes described here in both community composition and fermentation products due to ACA treatment, along with the statistical association between fecal SCFA concentrations and longevity, are consistent with this hypothesis, and provide a stepping-stone for future studies. Interestingly, SCFAs themselves have well-documented effects on glucose homeostasis (reviewed in [37] and [38]). The two explanations are therefore not mutually exclusive, and the effects of ACA on longevity may be mediated by both glucose physiology and microbial activity in the gut.

2.4 Conclusions

Here we have tested four predictions of a proposed model connecting ACA to lifespan via the gut microbiome. We demonstrate that ACA reproducibly modulates the composition of the microbiota, as well as the concentrations of fermentation products, increasing the abundance of butyrate and propionate. In addition, we provide evidence that the structure of the microbial community is an important factor in the composition of metabolites produced. Finally, we show an association between SCFA concentrations in feces and survival, suggesting a role of the microbiome in the life-extending properties of ACA. Together, these results encourage a new focus on managing the gut microbiota for host health and longevity.

2.5 Methods

2.5.1 Mouse housing and ACA treatment

All mice used in this study were maintained in specific-pathogen free conditions, and the protocols for husbandry and experimentation were approved by the Institutional Animal Care and Use Committees at each of the three institutions. Mice were bred and housed, and lifespan was assessed as described in [3]. Briefly, at each of the three study sites, genetically heterogeneous, UM-HET3, mice were produced by a four-way cross as previously described in [39]. After weaning, mice were fed LabDiet® (TestDiet Inc.) 5LG6 produced in common batches for all sites. At 42 days of age, electronic ID chips were surgically implanted and treatment randomly assigned to each cage housing four mice to a cage for females and three to a cage for males. ACA-treated mice were fed the same chow amended with 1,000 ppm ACA (Spectrum Chemical Manufacturing Corporation) from 8 months of age onwards. Mice were transferred

every 14 days to fresh, ventilated cages with water provided in bottles. Colonies at all three sites were assessed for infectious agents four times each year, and all tests were negative for the entire duration of the study.

2.5.2 Sample collection and processing

Fresh fecal pellets were collected directly from mice between 762 and 973 days of age and frozen at -80 . We did not control the time of day at collection. While differences in age and collection time could have added variability to SCFA concentrations, both were similarly distributed for the different treatment groups, so they are unlikely to confounded our analyses. To eliminate potential cage effects from co-housed mice [40], samples were obtained from no more than one randomly selected mouse per cage. A total of 144 samples were collected from 12 male and 12 female mice in both control and ACA treatment groups at each of the three sites. Samples were shipped on dry ice and then stored, frozen, until processing. For approximately the first half of samples, we extracted the soluble fraction by homogenizing pellets with 200 μL of nuclease-free water. For the remaining samples, we instead used a 1:10 ratio (weight:volume), with a maximum volume of 1.5 ml. This was found to improve quantification in higher weight samples. While SCFA concentration estimates were higher when using the amended protocol, the order of sample extraction was fully randomized, so it is unlikely to have biased our interpretations. Homogenized samples were centrifuged at 10,000 $\times$ g for 10 minutes to separate soluble and solid fractions. The supernatant was then serially vacuum filtered, ultimately through a 0.22 μm filter, before HPLC analysis. The solid fraction was frozen prior to DNA extraction. Four samples were excluded from chemical analysis and one from DNA analysis due to technical irregularities during sample processing.

Prior to DNA extraction, fecal pellet solids were thawed and, where necessary, subsampled for separate analysis. To move beyond relative abundance, solids were weighed and spiked with 10 μL aliquots of prepared *Sphingopyxis alaskensis* strain RB2256—an organism not found in mouse feces—in order to compare 16S rRNA gene abundance between samples [41, 42]. The spike was prepared as follows: a 1:200 dilution of a stationary phase *S. alaskensis* culture was grown at room temperature for approximately 44 hours in R2B medium with shaking. This culture was harvested at a final OD420 of 0.72 before being rinsed in PBS and resuspended—5-fold more concentrated—in 20% glycerol in PBS (v/v). Aliquots of these cells were stored at -20 before extraction and sequencing. Spiked fecal samples were homogenized in

nuclease free water at a ratio of 1:10 (w/v). DNA was extracted from 150 μL of this mixture using the MoBio PowerMag Microbiome kit.

2.5.3 Chemical analysis

The chemical composition of samples was assessed on a Shimadzu HPLC (Shimadzu Scientific Instruments) equipped with an RID-10A refractive index detector. 30 μL injections were run on an Aminex HPX-87H column (Bio-Rad Laboratories, Hercules, CA) at 50 with 0.01 N H_2SO_4 mobile phase and a flow rate of 0.6 ml/minute. External standards were run approximately daily containing acetate, butyrate, formate, glucose, lactate, propionate, and succinate at 8 concentrations between 0.1 mM to 20 mM. Due to the complexity of the chromatogram, the identity and area of retained peaks was curated manually, assisted by the LC Solutions Software (Shimadzu Scientific Instruments) Standard curves were fit using weighted regression (inverse square of the concentration), and, for all compounds except propionate, without an intercept.

2.5.4 16S rRNA gene sequencing and analysis

The V4 hypervariable region of the 16S rRNA gene was amplified from extracted DNA (as described in [43]), and sequenced on an Illumina MiSeq platform using a MiSeq Reagent Kit V2 500 cycles (cat# MS1022003). Amplicon sequences were processed with MOTHUR (version 1.39.4 [44]) using a protocol based on the 16S standard operating procedures [43]. Scripts to reproduce our analysis can be found at [45]. After fusing paired reads, quality trimming, and alignment to the SILVA reference database (Release 132 downloaded from [46]). The vast majority of 16S rRNA gene sequences were between 244 and 246 bp. Sequences were classified using the method of Wang *et al.* [47] as implemented in MOTHUR and with the SILVA non-redundant database as a reference [48]. We clustered sequences into OTUs using the OptiClust method [49] at a 97% similarity threshold. We counted and removed sequences classified as *S. alaskensis*, the spiked-in standard, before further analysis. We did not attempt to assess the exact number of 16S rRNA gene copies spiked into samples. Instead, spike-adjusted abundance was defined in units based on the standardized spike (μL spike equivalents / g sample) and estimated using the formula:

$$\text{RelativeAbundance} \times \frac{(\text{TotalEndemicReadCount} \times \text{SpikeVolume})}{(\text{TotalSpikeReadCount} \times \text{SampleWeight})}$$

Family level abundance was calculated as the summed abundance of all sequences clustered in OTUs classified to that family. OTU counts were randomly subsampled to the minimum number of reads before calculating Chao1 richness, but were not subsampled for other analyses. A single, independent realization of random subsampling was used for each richness calculation. A search of the NCBI non-redundant nucleotide database for related sequences from cultured bacteria was carried out using the BLASTn web tool [50] searching the non-redundant nucleotide database with default parameters and excluding sequences from uncultured organisms.

2.5.5 Statistical analysis

A 0.05 *P*-value threshold was used to define statistical significance, with values below 0.1 considered "suggestive". Except where specified, *P*-values are not corrected for multiple testing. Due to the risk of violating distributional assumptions, univariate comparisons between groups were done using the non-parametric Mann-Whitney U test. Differences in multivariate community composition and dispersion were tested using PERMANOVA (`adonis`) and PERMDISP (`betadisper`) respectively, both implemented in the vegan package (version 2.46 [51]) for the R programming language. Bray-Curtis dissimilarity was used as the β-diversity index.

Differences in the relative abundance of individual OTUs were surveyed using the DESeq2 package (version 1.18.1 [52]) for R, and fitting a model that included terms for treatment, sex, site, and the interaction between treatment and sex. So as to keep valuable distributional information, all OTUs found in at least two samples were included in the initial analysis, with *P*-values calculated using a Wald test. However, FDR correction using the Benjamini-Hochberg procedure [53] excluded "rare" OTUs—those with mean relative abundance less than 0.01% or detected in fewer than 5% of samples—in order to maintain statistical power by independently reducing the number of tests.

Interactions between sex and treatment in fecal SCFAs were assessed for log-transformed concentrations. The small number of zeros were replaced with half the lowest detected concentration for that metabolite. Interactions were tested in an ANOVA that also included terms for site, sex, and treatment. LASSO regressions of the three SCFAs and lactate against spike-adjusted OTU abundances were performed using the scikit-learn library for Python (version 0.18.2 [54]) and log-transformed concentrations after adjustment for site, sex, and treatment. OTUs detected in more than 5% of samples and with mean abundance greater than 0.01% were included. The

LASSO parameter was determined by randomized 10-fold cross-validation, optimizing for out-of-bag R^2. For each metabolite we confirmed that OTU abundance information improved predictions by testing the Spearman's rank correlation between true values and out-of-bag predictions of the best model using a Student's t-distribution approximation [55] and a $P = 0.05$ significance threshold. While this type of regularized regression is primarily useful for constructing predictive models, and biological interpretation can be challenging, non-zero regression coefficients are suggestive of covariates that are among the most strongly associated with a response. Proportional hazards regression was carried out using of the survival package (version 2.413 [56]) for R, and the day of fecal sampling as the entry time. All sampled mice were dead at the time of analysis and right-censoring was therefore not used. Standardized HRs reported for SCFAs are based on concentrations that have been centered around 0 and scaled to a standard deviation of 1.

2.5.6 Availability of data and materials

The sequence datasets generated and analyzed during the current study have been uploaded to the SRA database, accession SRP136736. Full-cohort survival data analyzed for portions of this study are available from the corresponding author on reasonable request. Code and metadata needed to reproduce the processing of raw data and downstream analyses is available on GitHub [45].

BIBLIOGRAPHY

[1] Smith, B.J., Miller, R.A., Ericsson, A.C., Harrison, D.E., Strong, R., Schmidt, T.M.: Changes in the gut microbiota and fermentation products associated with enhanced longevity in acarbose-treated mice. bioRxiv, 311456 (2018). doi:10.1101/311456

[2] Nadon, N.L., Strong, R., Miller, R.A., Harrison, D.E.: NIA Interventions Testing Program: Investigating Putative Aging Intervention Agents in a Genetically Heterogeneous Mouse Model. EBioMedicine **21**, 3–4 (2017). doi:10.1016/j.ebiom.2016.11.038

[3] Harrison, D.E., Strong, R., Allison, D.B., Ames, B.N., Astle, C.M., Atamna, H., Fernandez, E., Flurkey, K., Javors, M.A., Nadon, N.L., Nelson, J.F., Pletcher, S., Simpkins, J.W., Smith, D.L., Wilkinson, J.E., Miller, R.A.: Acarbose, 17-α-estradiol, and nordihydroguaiaretic acid extend mouse lifespan preferentially in males. Aging Cell **13**(2), 273–282 (2014). doi:10.1111/acel.12170

[4] Strong, R., Miller, R.A., Antebi, A., Astle, C.M., Bogue, M., Denzel, M.S., Fernandez, E., Flurkey, K., Hamilton, K.L., Lamming, D.W., Javors, M.A., de Magalhães, J.P., Martinez, P.A., McCord, J.M., Miller, B.F., Müller, M., Nelson, J.F., Ndukum, J., Rainger, G.E., Richardson, A., Sabatini, D.M., Salmon, A.B., Simpkins, J.W., Steegenga, W.T., Nadon, N.L., Harrison, D.E.: Longer lifespan in male mice treated with a weakly estrogenic agonist, an antioxidant, an α-glucosidase inhibitor or a Nrf2-inducer. Aging Cell **15**(5), 872–884 (2016). doi:10.1111/acel.12496

[5] Laube, H.: Acarbose: An update of its therapeutic use in diabetes treatment. Clinical Drug Investigation **22**(3), 141–156 (2002)

[6] Hanefeld, M., Cagatay, M., Petrowitsch, T., Neuser, D., Petzinna, D., Rupp, M.: Acarbose reduces the risk for myocardial infarction in type 2 diabetic patients: Meta-analysis of seven long-term studies. European Heart Journal **25**(1), 10–16 (2004). doi:10.1016/S0195-668X(03)00468-8

[7] Dehghan-Kooshkghazi, M., Mathers, J.C.: Starch digestion, large-bowel fermentation and intestinal mucosal cell proliferation in rats treated with the α-glucosidase inhibitor acarbose. British Journal of Nutrition **91**(03), 357 (2004). doi:10.1079/BJN20031063

[8] Weaver, G.A., Tangel, C.T., Krause, J.A., Parfitt, M.M., Jenkins, P.L., Rader, J.M., Lewis, B.A., Miller, T.L., Wolin, M.J.: Acarbose Enhances Human Colonic Butyrate Production. The Journal of Nutrition **127**(5), 717–723 (1997). doi:10.1093/jn/127.5.717

[9] Hiele, M., Ghoos, Y., Rutgeerts, P., Vantrappen, G.: Effects of acarbose on starch hydrolysis. Digestive Diseases and Sciences **37**(7), 1057–1064 (1992). doi:10.1007/BF01300287

[10] Seifarth, C., Bergmann, J., Holst, J.J., Ritzel, R., Schmiegel, W., Nauck, M.A.: Prolonged and enhanced secretion of glucagon-like peptide 1 (7-36 amide) after oral sucrose due to α-glucosidase inhibition (acarbose) in Type 2 diabetic patients. Diabetic Medicine **15**(6), 485–491 (1998). doi:10.1002/(SICI)1096-9136(199806)15:6<485::AID-DIA610>3.0.CO;2-Y

[11] Qualmann, C., Nauck, M.A., Holst, J.J., Orskov, C., Creutzfeldt, W.: Glucagon-like peptide 1 (7-36 amide) secretion in response to luminal sucrose from the upper and lower gut. A study using α-glucosidase inhibition (acarbose). Scandinavian Journal of Gastroenterology **30**(9), 892–6 (1995). doi:10.3109/00365529509101597

[12] Jenkins, D.J.A., Taylor, R.H., Goff, D.V., Fielden, H., Misiewicz, J.J., Sarson, D.L., Bloom, S.R., Alberti, K.G.M.M.: Scope and specificity of acarbose in slowing carbohydrate absorption in man. Diabetes **30**(11), 951–954 (1981). doi:10.2337/DIAB.30.11.951

[13] Weaver, G.A., Tangel, C.T., Krause, J.A., Parfitt, M.M., Stragand, J.J., Jenkins, P.L., Erb, T.A., Davidson, R.H., Alpern, H.D., Guiney, W.B., Higgins, P.J.: Biomarkers of human colonic cell growth are influenced differently by a history of colonic neoplasia and the consumption of acarbose. The Journal of Nutrition **130**(11), 2718–25 (2000). doi:10.1093/jn/130.11.2718

[14] Holt, P.R., Atillasoy, E., Lindenbaum, J., Ho, S.B., Lupton, J.R., McMahon, D., Moss, S.F.: Effects of acarbose on fecal nutrients, colonic pH, and short-chain fatty acids and rectal proliferative indices. Metabolism: Clinical and Experimental **45**(9), 1179–1187 (1996). doi:10.1016/S0026-0495(96)90020-7

[15] Wolever, T.M.S., Chiasson, J.L.: Acarbose raises serum butyrate in human subjects with impaired glucose tolerance. British Journal of Nutrition **84**(1), 57–61 (2000). doi:10.1017/S0007114500001239

[16] Wolin, M.J., Miller, T.L., Yerry, S., Bank, S., Weaver, G.A., Zhang, Y.: Changes of Fermentation Pathways of Fecal Microbial Communities Associated with a Drug Treatment That Increases Dietary Starch in the Human Colon Changes of Fermentation Pathways of Fecal Microbial Communities Associated with a Drug Treatment That Increas. Applied and Environmental Microbiology **65**(7), 2807–2812 (1999)

[17] Zhang, X., Fang, Z., Zhang, C., Xia, H., Jie, Z., Han, X., Chen, Y., Ji, L.: Effects of Acarbose on the Gut Microbiota of Prediabetic Patients: A Randomized, Double-blind, Controlled Crossover Trial. Diabetes Therapy **8**(2), 293–307 (2017). doi:10.1007/s13300-017-0226-y

[18] Zhao, L., Zhang, F., Ding, X., Wu, G., Lam, Y.Y., Shi, Y., Shen, Q., Dong, W., Liu, R., Ling, Y., Zeng, Y.: Gut bacteria selectively promoted by dietary fibers alleviate type 2 diabetes. Science **1156**(March), 1151–1156 (2018). doi:10.1126/science.aao5774

[19] Koh, A., De Vadder, F., Kovatcheva-Datchary, P., Bäckhed, F.: From dietary fiber to host physiology: Short-chain fatty acids as key bacterial metabolites. Cell **165**(6), 1332–1345 (2016). doi:10.1016/j.cell.2016.05.041

[20] Kasubuchi, M., Hasegawa, S., Hiramatsu, T., Ichimura, A., Kimura, I.: Dietary gut microbial metabolites, short-chain fatty acids, and host metabolic regulation. Nutrients **7**(4), 2839–2849 (2015). doi:10.3390/nu7042839

[21] Wolever, T.M.S., Fernandes, J., Rao, A.V.: Serum Acetate:Propionate Ratio Is Related to Serum Cholesterol in Men but Not Women. The Journal of Nutrition **126**(11), 2790–2797 (1996). doi:10.1093/jn/126.11.2790

[22] Lowe, P.P., Gyongyosi, B., Satishchandran, A., Iracheta-Vellve, A., Ambade, A., Kodys, K., Catalano, D., Ward, D.V., Szabo, G.: Alcohol-related changes in the intestinal icrobiome influence neutrophil infiltration, inflammation and

steatosis in early alcoholic hepatitis in mice. PLoS ONE **12**(3), 1–16 (2017). doi:10.1371/journal.pone.0174544

[23] Castoldi, A., Andrade-Oliveira, V., Aguiar, C.F., Amano, M.T., Lee, J., Miyagi, M.T., Latância, M.T., Braga, T.T., da Silva, M.B., Ignácio, A., Carola Correia Lima, J.D., Loures, F.V., Albuquerque, J.A.T., Macêdo, M.B., Almeida, R.R., Gaiarsa, J.W., Luévano-Martínez, L.A., Belchior, T., Hiyane, M.I., Brown, G.D., Mori, M.A., Hoffmann, C., Seelaender, M., Festuccia, W.T., Moraes-Vieira, P.M., Câmara, N.O.S.: Dectin-1 Activation Exacerbates Obesity and Insulin Resistance in the Absence of MyD88. Cell Reports **19**(11), 2272–2288 (2017). doi:10.1016/j.celrep.2017.05.059

[24] Singer, B.H., Dickson, R.P., Denstaedt, S.J., Newstead, M.W., Kim, K., Falkowski, N.R., Erb-Downward, J.R., Schmidt, T.M., Huffnagle, G.B., Standiford, T.J.: Bacterial Dissemination to the Brain in Sepsis. American Journal of Respiratory and Critical Care Medicine **197**(6), 747–756 (2018). doi:10.1164/rccm.201708-1559OC

[25] Ormerod, K.L., Wood, D.L.A., Lachner, N., Gellatly, S.L., Daly, J.N., Parsons, J.D., Dal'Molin, C.G.O., Palfreyman, R.W., Nielsen, L.K., Cooper, M.A., Morrison, M., Hansbro, P.M., Hugenholtz, P.: Genomic characterization of the uncultured Bacteroidales family S24-7 inhabiting the guts of homeothermic animals. Microbiome **4**(1), 36 (2016). doi:10.1186/s40168-016-0181-2

[26] Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R., Tiedje, J.M.: Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Research **42**(Database issue), 633–642 (2014). doi:10.1093/nar/gkt1244

[27] Clavel, T., Lagkouvardos, I., Blaut, M., Stecher, B.: The mouse gut microbiome revisited: From complex diversity to model ecosystems. International Journal of Medical Microbiology **306**(5), 316–327 (2016). doi:10.1016/j.ijmm.2016.03.002

[28] Salzman, N.H., de Jong, H., Paterson, Y., Harmsen, H.J.M., Welling, G.W., Bos, N.A.: Analysis of 16S libraries of mouse gastrointestinal microflora reveals a large new group of mouse intestinal bacteria. Microbiology **148**(11), 3651–3660 (2002). doi:10.1099/00221287-148-11-3651

[29] Lagkouvardos, I., Pukall, R., Abt, B., Foesel, B.U., Meier-Kolthoff, J.P., Kumar, N., Bresciani, A., Martínez, I., Just, S., Ziegler, C., Brugiroux, S., Garzetti, D., Wenning, M., Bui, T.P.N., Wang, J., Hugenholtz, F., Plugge, C.M., Peterson, D.A., Hornef, M.W., Baines, J.F., Smidt, H., Walter, J., Kristiansen, K., Nielsen, H.B., Haller, D., Overmann, J., Stecher, B., Clavel, T.: The Mouse Intestinal Bacterial Collection (miBC) provides host-specific insight into cultured diversity and functional potential of the gut microbiota. Nature Microbiology **1**(August), 16131 (2016). doi:10.1038/nmicrobiol.2016.131

[30] Wexler, H.M.: Bacteroides: The good, the bad, and the nitty-gritty. Clinical Microbiology Reviews **20**(4), 593–621 (2007). doi:10.1128/CMR.00008-07

[31] Song, Y., Liu, C., Finegold, S.M.: Bacteroides. In: Whitman, W.B. (ed.) Bergey's Manual of Systematics of Archaea and Bacteria. John Wiley & Sons, Ltd, Hoboken, NJ (2015). doi:10.1002/9781118960608.gbm00238

[32] Macy, J.M., Ljungdahl, L.G., Gottschalk, G.: Pathway of succinate and propionate formation in Bacteroides fragilis. Journal of Bacteriology **134**(1), 84–91 (1978)

[33] Macfarlane, S., Macfarlane, G.T.: Regulation of short-chain fatty acid production. Proceedings of the Nutrition Society **62**(1), 67–72 (2003). doi:10.1079/PNS2002207

[34] Weaver, G.A., Krause, J.A., Miller, T.L., Wolin, M.J.: Cornstarch fermentation by the colonic microbial community yields more butyrate than does cabbage fiber fermentation; cornstarch fermentation rates correlate negatively with methanogenesis. American Journal of Clinical Nutrition **55**(1), 70–77 (1992). doi:10.1093/ajcn/55.1.70

[35] Cummings, J.H., Englyst, H.N.: Fermentation in the human large intestine and the available substrates. American Journal of Clinical Nutrition **45**, 1243–1255 (1987)

[36] Miller, R.A., Harrison, D.E., Astle, C.M., Fernandez, E., Flurkey, K., Han, M., Javors, M.A., Li, X., Nadon, N.L., Nelson, J.F., Pletcher, S., Salmon, A.B., Sharp, Z.D., Van Roekel, S., Winkleman, L., Strong, R.: Rapamycin-mediated lifespan increase in mice is dose and sex dependent and metabolically distinct from dietary restriction. Aging Cell **13**(3), 468–477 (2014). doi:10.1111/acel.12194

[37] Morrison, D.J., Preston, T.: Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism. Gut Microbes **7**(3), 189–200 (2016). doi:10.1080/19490976.2015.1134082

[38] Canfora, E.E., Jocken, J.W., Blaak, E.E.: Short-chain fatty acids in control of body weight and insulin sensitivity. Nature Reviews Endocrinology **11**(10), 577–591 (2015). doi:10.1038/nrendo.2015.128

[39] Miller, R.A., Harrison, D.E., Astle, C.M., Baur, J.A., Boyd, A.R., de Cabo, R., Fernandez, E., Flurkey, K., Javors, M.a., Nelson, J.F., Orihuela, C.J., Pletcher, S., Sharp, Z.D., Sinclair, D.A., Starnes, J.W., Wilkinson, J.E., Nadon, N.L., Strong, R.: Rapamycin, but not resveratrol or simvastatin, extends life span of genetically heterogeneous mice. Journals of Gerontology - Series A Biological Sciences and Medical Sciences **66 A**(2), 191–201 (2011). doi:10.1093/gerona/glq178

[40] Laukens, D., Brinkman, B.M., Raes, J., De Vos, M., Vandenabeele, P.: Heterogeneity of the gut microbiome in mice: Guidelines for optimizing experimental design. FEMS Microbiology Reviews **40**(1), 117–132 (2015). doi:10.1093/femsre/fuv036

[41] Smets, W., Leff, J.W., Bradford, M.A., McCulley, R.L., Lebeer, S., Fierer, N.: A method for simultaneous measurement of soil bacterial abundances and community composition via 16S rRNA gene sequencing. Soil Biology and Biochemistry **96**, 145–151 (2016). doi:10.1016/j.soilbio.2016.02.003

[42] Stämmler, F., Gläsner, J., Hiergeist, A., Holler, E., Weber, D., Oefner, P.J., Gessner, A., Spang, R.: Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. Microbiome, 1–13 (2016). doi:10.1186/s40168-016-0175-0

[43] Kozich, J.J., Westcott, S.L., Baxter, N.T., Highlander, S.K., Schloss, P.D.: Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the miseq illumina sequencing platform. Applied and Environmental Microbiology **79**(17), 5112–5120 (2013). doi:10.1128/AEM.01043-13

[44] Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., Sahl, J.W., Stres, B., Thallinger, G.G., Van Horn, D.J., Weber, C.F.: Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology **75**(23), 7537–7541 (2009). doi:10.1128/AEM.01541-09

[45] Smith, B.J.: Code and Metadata to Reproduce: Changes in the gut microbiota and fermentation products associated with enhanced longevity in acarbose-treated mice. (2018). doi:10.5281/zenodo.1229203. https://github.com/bsmith89/smith2018paper/releases/tag/v0.1 Accessed 2018-04-25

[46] Schloss, P.D.: Silva reference files (2018). https://www.mothur.org/wiki/Silva{_}reference{_}files Accessed 2018-02-01

[47] Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R.: Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology **73**(16), 5261–5267 (2007). doi:10.1128/AEM.00062-07. Wang, Qiong, 2007, Naive

[48] Yilmaz, P., Parfrey, L.W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., Schweer, T., Peplies, J., Ludwig, W., Glöckner, F.O.: The SILVA and "all-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Research **42**(D1), 643–648 (2014). doi:10.1093/nar/gkt1209

[49] Westcott, S.L., Schloss, P.D.: OptiClust, an Improved Method for Assigning Amplicon-Based Sequence Data to Operational Taxonomic Units. mSphere **2**(2), 00073–17 (2017). doi:10.1128/mSphereDirect.00073-17

[50] Wheeler, D.L., Chappey, C., Lash, A.E., Leipe, D.D., Madden, T.L., Shuler, G.D., Tatusova, T.A., Rapp, B.A.: Database resources of the National Center for Biotechnology Information. Nucleic Acids Research **41**(November 2012), 8–20 (2000). doi:10.1093/nar/gks1189

[51] Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E., Wagner, H.: vegan: Community Ecology Package (2018). https://cran.r-project.org/package=vegan

[52] Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology **15**(12), 550 (2014). doi:10.1186/s13059-014-0550-8

[53] Benjamini, Y., Hochberg, Y.: Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **57**(1), 289–300 (1995)

[54] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011). 1201.0490

[55] Iman, R.L., Conover, W.J.: Approximations of the critical region for spearman's rho with and without ties present. Communications in Statistics - Simulation and Computation **7**(3), 269–282 (1978). doi:10.1080/03610917808812076

[56] Therneau, T.M.: A Package for Survival Analysis in S (2015). https://cran.r-project.org/package=survival

CHAPTER 3

Muribaculaceae genomes assembled from metagenomes suggest genetic drivers of differential response to acarbose treatment in mice

3.1 Background

The mammalian gut microbiome is a complex ecological system that influences energy balance [1], pathogen resistance [2], and inflammation [3], among other processes with importance to host health. Understanding how the bacterial inhabitants of the gut respond to pharmaceutical and dietary perturbations is a major step in developing a predictive framework for microbiome-based therapies. Acarbose (ACA) is an α-glucosidase inhibitor prescribed for the treatment of type 2 diabetes mellitus because it reduces the absorption of glucose from starch in the small intestine [4]. In rodents, ACA has been shown to increase the amount of starch entering the lower digestive system after a meal [5], resulting in changes to the composition of the gut microbiota and its fermentation products [5–12]. Interestingly, long-term treatment with ACA has been shown to substantially increase longevity in male mice and to a lesser extent in females [13, 14].

In Chapter 2 it was shown that the relative abundance of a number of bacterial taxa as well as the concentrations of propionate and butyrate respond to long term treatment with ACA. This study was notable in being replicated across three sites: The University of Michigan (UM) in Ann Arbor, The University of Texas Health Science Center at San Antonio (UT), and The Jackson Laboratory (TJL) in Bar Harbor, Maine. At UM and TJL one highly abundant operational taxonomic unit (OTU), classified as a member of the *Bacteroidales* family *Muribaculaceae* and here desig-

nated as B1, was found to be enriched nearly 4-fold in ACA treated mice. B1 was also present and abundant at UT but was not found to be significantly more abundant in ACA treated mice relative to controls. Instead, a different member of the *Muribaculaceae*, designated B2, was found to be highly abundant and 4-fold enriched in ACA-treated mice, but was nearly absent at UM and TJL. Other *Muribaculaceae* were also identified as among the most abundant members of the mouse gut microbiota across the three sites, although none of these were found to be enriched in ACA treatment.

Family *Muribaculaceae*—formerly the S24-7 and sometimes referred to as *Candidatus Homeothermaceae*—has only one published cultivar [15] despite being a common and abundant inhabitant of the mammalian gut, especially in mice [16]. Previous studies have suggested that the *Muribaculaceae* specialize on the fermentation of complex polysaccharides [16], much like members of the genus *Bacteroides* also in order *Bacteroidales*.

Recently, techniques have been developed for the reconstruction of genomes of uncultivated members of bacterial communities [17, 18]. Based on 30 such metagenome assembled genomes (MAGs) they reconstructed using this approach, Ormerod *et al.* [16] proposed that the *Muribaculaceae* fall into three distinct carbohydrate utilization guilds, which they describe as specialists on α-glucans, plant glycans, and host glycans, respectively. While it is reasonable to expect that α-glucan specialists would be most benefited by the large influx of starch to the gut resulting from ACA treatment, this prediction has not been tested, and physiological inferences based on the genome content of members of this clade have been largely divorced from biological observations.

Experimental perturbations of complex microbial communities present an opportunity to observe ecological features of many bacterial taxa without cultivated members and generate hypotheses about their physiology. Given the observed, dramatically increased relative abundance of B1 and B2 (here referred to as "responders") in mice treated with ACA, we hypothesize that these OTUs are capable of robust growth on starch, while the other *Muribaculaceae* found in the study ("non-responders"), lack the genomic features necessary for the utilization of the polysaccharide. Alternatively, responders may be resistant to the inhibitory effects of ACA, or benefit from elevated levels of intermediate starch degradation products. Since isolates of the *Muribaculaceae* species in these mice are not available for characterization, a comparative genomic approach is taken to explore their functional potential.

Most of the research on the genomic components of polysaccharide degradation

in gram negative bacteria has been carried out in the genus *Bacteroides*, and in particular *B. thetaiotaomicron* [19]. Starch utilization in *B. thetaiotaomicron* is dependent on an ensemble of eight proteins, SusRABCDEFG that enable recognition, binding, hydrolysis, and import of starch and related polysaccharides [20]. Homologs of SusC and SusD characterize all known polysaccharide utilization systems in this clade [21], are encoded in Sus-like genomic regions known as polysaccharide utilization loci (PULs), and are widespread in the *Bacteroidetes* [22]. The molecular range of these systems is determined by the carbohydrate-active enzymes and structural proteins they encode, based on the specificity of glycoside hydrolase (GH) and carbohydrate binding module (CBM) domains, which have been extensively cataloged in the dbCAN database [23, 24].

Here MAGs from the feces of mice at UT and UM are analyzed to explore two closely related questions about the niche of B1 and B2 in the lower digestive system. First, why do B1 and B2 each increase with ACA treatment, while other *Muribaculaceae* do not? And second, why is the response of B1 site specific? Despite similar patterns of abundance at their respective sites, these two OTUs seem to be only distantly related, sharing just 90% of nucleotides in their 16S rRNA gene V4 hypervariable region (see Appendix A). We nonetheless find genomic evidence that B1 and B2 occupy overlapping niches, specializing in the degradation of α-glucans, a role not held by the other *Muribaculaceae* described in this study. In addition, we identify two distinct variants of B1, referred to as B1-A and B1-B, which are differentially distributed between UM and UT and have functionally relevant differences in gene content.

Reconstructing genomes from metagenomes allow for the comparison of the functional potential of *Muribaculaceae* at UM and UT. This work demonstrates the utility of culture-free genomics to understand the ecological role of these key members of the mouse gut microbial community and explore several hypotheses that may explain differences in the distribution and response of bacteria to perturbations. Hypotheses derived from this analysis provide a foundation for future physiological studies in recently obtained cultivars. While a preponderance of host-associated bacterial species have never been isolated, let alone characterized [25], combining experimental data from complex communities with the analysis of reconstructed genomes provides a powerful tool for expanding understanding to these understudied taxa.

3.2 Results

3.2.1 Recovered population genomes are of high quality and resemble other *Muribaculaceae* genomes

MAGs were constructed for 7 populations in the family *Muribaculaceae*, including ACA responders B1 and B2, and non-responders B3 through B7. For B1, two genomic variants were recovered, B1-A and B1-B, MAGs that possess 0.63 and 0.36 Mbp of unshared sequence, respectively (additional details about these variants are in Section 3.2.3). All 8 novel MAGs are estimated to be of high completeness and all had less than 1% estimated contamination based on the recovery of ubiquitous, single-copy genes. The median N50 statistic was approximately 71 kbp, indicating successful assembly, and suggesting that inferences based on genomic context are generally possible. Estimated genome sizes, GC%, and number of predicted genes are all similar to previously published MAGs as well as the finished *Muribaculum intestinale* YL27 genome.

Table 3.1: Summary of novel MAGs compared to the genome of *Muribaculum intestinale* YL27

Taxon	Completeness[1]	Scaffolds	Length[2]	N50	GC	in Chapter 2
YL-27[3]	99%	1	3.3	3,307,069	50.1%	
B1-A	97%	228	3.2	41,412	46.6%	OTU-1
B1-B	97%	152	3.0	59,916	46.9%	OTU-1
B2	98%	65	2.6	79,454	50.5%	OTU-4
B3	86%	98	2.2	63,818	54.0%	OTU-6
B4	98%	31	2.7	148,039	55.2%	OTU-5
B5	86%	50	2.5	78,179	55.7%	OTU-8
B6	99%	110	3.2	87,115	48.3%	OTU-30
B7	98%	97	2.5	59,037	53.9%	OTU-39

[1] Estimated by CheckM [26]
[2] Total length in Mbp
[3] *Muribaculum intestinale* YL-27 reference genome

In order to confirm the assertion that each of the reconstructed genomes is representative of *Muribaculaceae* OTU described in Chapter 2, per library mapping rates of each genome were compared to the relative abundance of the associated 16S rRNA gene in amplicon libraries. Despite the biases and technical variability inherent to both sequencing methods, and the limitations of mapping software, Pearson correlation coefficients between the fraction of reads mapped and OTU relative abundance

were above 0.86 for all MAGs,

3.2.1.1 Phylogenetics

To better understand the evolutionary relationships between these organisms, a concatenated gene tree was constructed for all 8 novel MAGs, as well as 30 publicly available MAG sequences [16], and *M. intestinale* YL27. The tree was rooted by four other *Bacteroidales* species: *Bacteroides ovatus* (ATCC-8483), *Bacteroides thetaiotaomicron* VPI-5482, *Porphyromonas gingivalis* (ATCC-33277), and *Barnesiella viscericola* (DSM-18177). Most internal nodes were found to have high topological confidence, and the placement of the MAGs reconstructed by Ormerod *et al.* was highly consistent with their published tree. To check that this concatenated approach is reflective of the organismal evolutionary history, a second maximum likelihood tree was constructed based on the *rpoB* gene, which is generally not thought to be transmitted horizontally, (despite exceptions [27]), also recapitulating the published topology. The estimated phylogeny shows that the 8 OTUs with newly reconstructed MAGs encompass most of the documented diversity of *Muribaculaceae*. Two of our taxa, B2 and B6, appear to be closely related to taxa with genomes reconstructed by Ormerod *et al.*: M6, and M1, respectively. Nonetheless, this phylogenetic analysis suggests that many of the genomes reconstructed here have not been described previously.

3.2.1.2 Novel protein families

Annotations based on alignment to a database of previously characterized sequences may provide only limited insight, in particular for genomes from largely unstudied families of bacteria. In order to identify previously uncharacterized orthologous groups, *de novo* clustering [28] was carried out based on amino acid similarity of all putative genes found in the 8 novel MAGs, 30 previously reconstructed MAGs, *M. intestinale*, four publicly available draft genomes from the family, and the four reference *Bacteroidales*. The resulting clusters are referred to as operational protein families (OPFs). While a fraction of the 12,648 resulting OPFs may be due to spurious sequence similarity and without biological relevance, 5,767 had representatives in at least three genomes, increasing the likelihood that these reflect evolutionarily conserved protein sequences. Of these, only 2,404 had members annotated with any COG, KO, or putative function. The remaining 3,363 OPFs include 17,831 predicted proteins across the 47 genomes

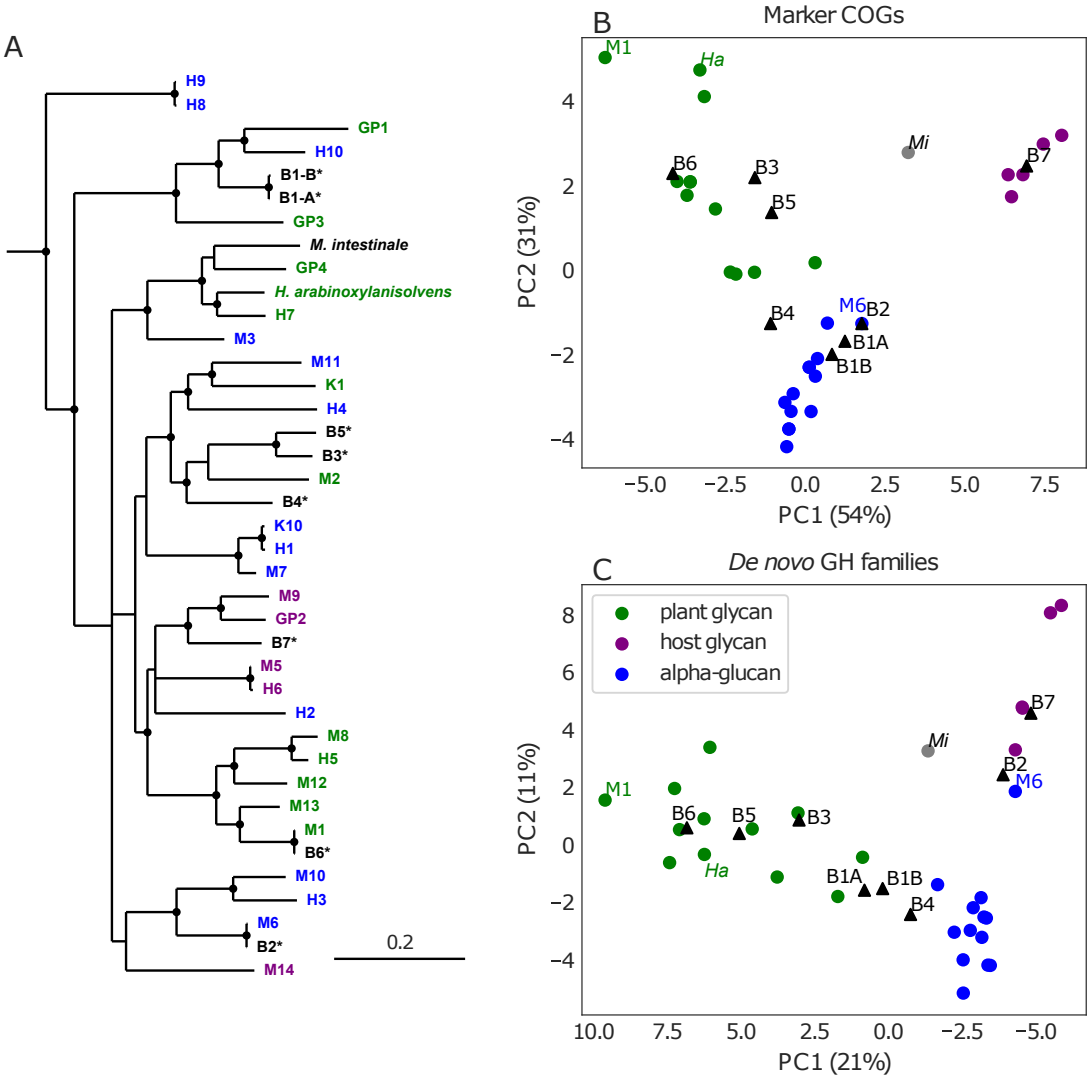
A
H9
H8
GP1
H10
B1-B*
B1-A*
GP3
M. intestinale
GP4
H. arabinoxylanisolvens
H7
M3
M11
K1
H4
B5*
B3*
M2
B4*
K10
H1
M7
M9
GP2
B7*
M5
H6
H2
M8
H5
M12
M13
M1
B6*
M10
H3
M6
B2*
M14
0.2
B
Marker COGs
PC2 (31%)
PC1 (54%)
C
De novo GH families
plant glycan
host glycan
alpha-glucan
PC2 (11%)
PC1 (21%)

Figure 3.1: Comparison of novel and previously described *Muribaculaceae* genomes. Novel MAGs (for OTUs "B1" through "B7") are combined with the finished genome for *M. intestinale* strain YL27, as well as 30 MAGs reconstructed by Ormerod *et al.*, hypothesized to reflect three polysaccharide utilization guilds: specializing on α-glucans (points and labels colored blue), host glycans (violet), and plant glycans (green). (**A**) Novel MAGs were placed in a phylogenetic context using a maximum-likelihood concatenated gene tree based on an amino-acid alignment of 9 shared, single-copy genes, and four other *Bacteroidales* species as an outgroup. Nodes with less than 70% confidence are collapsed into polytomies and topological support greater than 95% is indicated (black dots). Branch length indicates an estimate of expected substitutions per site. (**B, C**) Functional comparisons were visualized by plotting the first two principal components of an ordination on annotation counts of either (**B**) eight COGs identified by Ormerod *et al.* as maximally discriminatory between hypothesized guilds, or (**C**) *de novo* clusters based on sequence similarity of GH domain containing proteins. PCA was performed on the 30 MAGs reconstructed by Ormerod *et al.* and the percent of variation described by the first two components is included in the axis labels. All genomes were then projected onto that space. Novel MAGs (black triangles) are labeled, as are the previously described MAGs M1, M6, and the proposed *H. arabinoxylanisolvens* (Ha), and the finished genome of *M. intestinale* (Mi, grey circle).

3.2.1.3 Annotation ordination

To compare novel MAGs to other available genomes, a previous published analysis was recreated, harnessing a set of 8 COGs found by Ormerod *et al.* to maximally differentiate the three hypothesized guilds. By projecting genome annotations onto a reproduction of this previously defined space (see Figure 3.1), newly available genomes were compared to the three clusters hypothesized to represent specialization on α-glucans, plant glycans, and host glycans. While the 8 novel MAGs inhabit approximately the same volume as those previously reconstructed, and some could be plausibly classified based on these criteria, the ambiguous placement of B4 and *M. intestinale* suggests that new genomes will present additional exceptions to the three-guild model.

It is notably that both responders cluster with the proposed α-glucan guild, consistent with a functional potential for starch utilization not present in the non-responders. To expand on this descriptive analysis and to leverage the more comprehensive view provided by *de novo* clustering to explore differences and similarities in carbohydrate utilization potential, a second ordination of genomes was performed, this time based on OPF labels of predicted genes found to contain GH domains (Figure 3.2). Similar to the previous ordination based on COGs, three groups of genomes approximately reflecting those proposed by Ormerod *et al.* are apparent. However, the placement of B2 (as well as the closely related M6) relative to the proposed guilds are substantially different.

3.2.2 Comparison of responder and non-responder MAGs suggest genomic features with role in starch utilization

Based on the characterization of genes and genomic regions with a role in starch utilization in the closely related genus *Bacteroides*, it is plausible that α-amylase localized to the outer membrane may be common to starch utilizing bacteria in the order *Bacteroidales* [29]. Indeed, B1 has three OM-localized genes predicted to code for GH13 containing lipoproteins (B1A_280, B1A_301, B1A_333), each in a separate PUL (see Figure 3.2). While it also includes members without this activity, GH13 is the main family of α-amylases [30]. These genomic regions also possess additional genes with carbohydrate-active domains that are expected to interact with α-glucans.

Besides B1, B5 is the only other OTU to possess a putative PUL coding for a full complement of predicted starch-active proteins. Several OPFs have members in both this region and either B1 or *B. thetaiotaomicron* PULs, suggesting shared

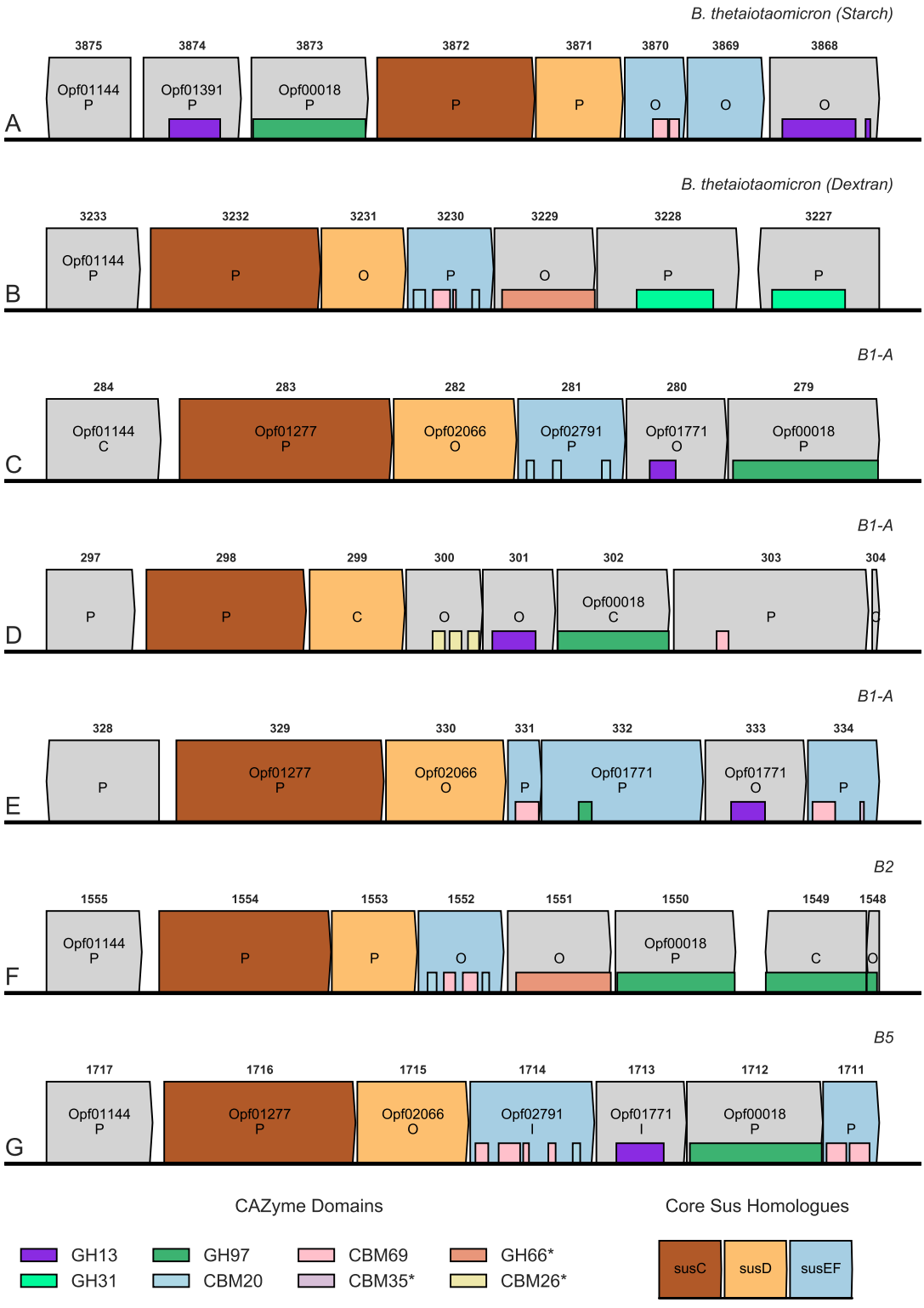
B. thetaiotaomicron (Starch)
A
3875 3874 3873 3872 3871 3870 3869 3868
Opf01144 P
Opf01391 P
Opf00018 P
P
P
O
O
O
B. thetaiotaomicron (Dextran)
B
3233 3232 3231 3230 3229 3228 3227
Opf01144 P
P
O
P
O
P
P
B1-A
C
284 283 282 281 280 279
Opf01144 C
Opf01277 P
Opf02066 O
Opf02791 P
Opf01771 O
Opf00018 P
B1-A
D
297 298 299 300 301 302 303 304
P
P
C
O
O
Opf00018 C
P
C
B1-A
E
328 329 330 331 332 333 334
P
Opf01277 P
Opf02066 O
P
Opf01771 P
Opf01771 O
P
B2
F
1555 1554 1553 1552 1551 1550 1549 1548
Opf01144 P
P
P
O
O
Opf00018 P
C
O
B5
G
1717 1716 1715 1714 1713 1712 1711
Opf01144 P
Opf01277 P
Opf02066 O
Opf02791 I
Opf01771 I
Opf00018 P
P
CAZyme Domains
GH13
GH31
GH97
CBM20
CBM69
CBM35*
GH66*
CBM26*
Core Sus Homologues
susC
susD
susEF

Figure 3.2: Polysaccharide utilization loci in *Bacteroidales*. Diagrams of the Sus operon (**A**) and the dextran associated PUL (**B**) of *B. thetaiotaomicron* along with five putative starch-associated PULs identified in three *Muribaculaceae* MAGs (**C-G**). Predicted protein coding sequences are shown as boxes pointed in the direction of transcription. Homology to SusC, SusD, and SusEF is indicated. Protein regions with homology to starch-associated GHs, as well as GH66, and CBMs are shown as shallow rectangles, and are colored as indicated in the legend. Several OPFs are noted with members in multiple genomes, including clusters that contain SusR (Opf01144), SusA (Opf01391), and SusB (Opf00018). The inferred localization of each protein product is also indicated: cytoplasmic (genes labeled C), periplasmic (P), outer membrane (O), or inner membrane (I).

function. This set including SusC-homologs Opf01277, Opf02066, which includes relatives of SusD, and Opf02791 whose members possess CBM20 starch-binding domains. However, while B5 also has a GH13 containing lipoprotein (B5_1713), its predicted localization is on the inner membrane. It is unclear whether this explains B5's non-response in ACA-treated mice. Plausible OM-localized, GH13 containing proteins are not found in any non-responders. While this characteristic does not seem to perfectly discriminate responder from non-responder OTUs—B2 also lacks such a gene—it nonetheless demonstrates concordance between inferred genomic features and observed population dynamics.

Despite the absence of a GH13 domain on the outer-membrane, it is plausible that B2 is capable of degrading starch using other enzymatic machinery. We speculate about one putative locus (see Figure 3.2 panel F), which has a similar gene content to characterized [31–33] dextran PULs in *B. thetaiotaomicron* and *B. ovatus*.

To expand the search for relevant genetic features, *de novo* protein clusters were filtered to those with members in the MAGs for both B1 and B2. Of these OPFs, several stood out as particularly relevant. Opf01144 includes SusR, the regulator of transcription of the starch utilization system in *B. thetaiotaomicron*, as well as its homolog in *B. ovatus*. It is an apparent subcluster of the larger family defined by K21557, and in many cases is encoded directly upstream of *susC* in putative PULs which consider likely to have affinity for α-glucans. In B1, two of the three putative starch PULs encode a member of Opf01144, and it is similarly located in PULs with starch-active CBM and GH domains in B2 and B5. In addition, of the seven MAGs reconstructed by Ormerod *et al.* that encode a member of this cluster, five of them are classified to the α-glucan guild. It is plausible that members of Opf01144 share a functional role regulating transcriptional responses to α-glucans.

Opf01391, which recapitulates K21575, includes SusA: the periplasmic neopullulanase of *B. thetaiotaomicron* and an important component of starch utilization in that organism [34]. This family is found in the MAGs of both responders, B1 and B2, and none of the non-responders. What's more, it's found in twelve of the thirteen α-glucan and a minority of the plant glycan guild members. Interestingly, although it is encoded by the Sus operon in *B. thetaiotaomicron* and its homologous locus in *B. ovatus*, in the *Muribaculaceae* members of Opf01391 do not in general appear to be encoded in PULs.

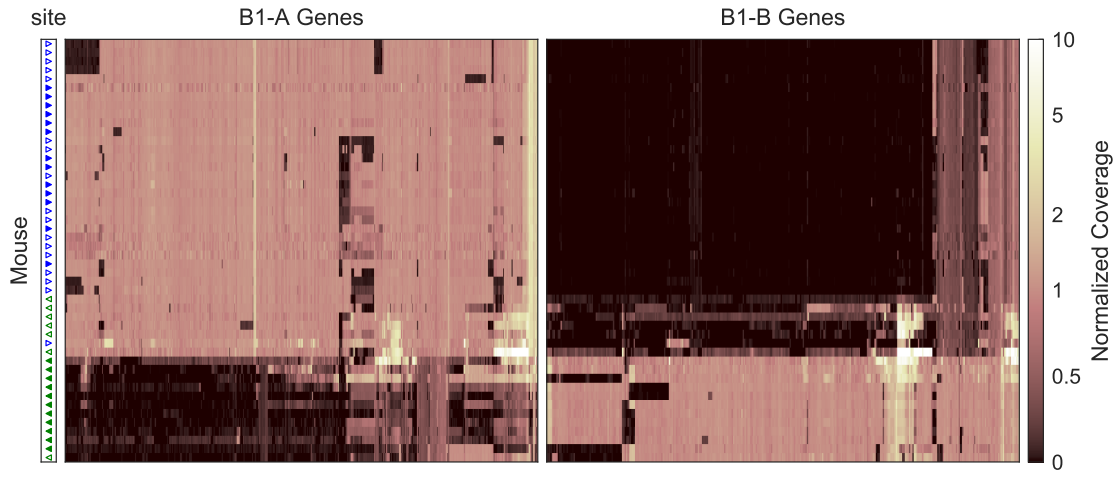

Figure 3.3: Visualization of differential gene content in two B1 populations. Heatmaps depict mapping coverage of metagenomes against putative protein coding genes in the B1-A or B1-B MAG normalized to the median coverage. Rows represent one or more pooled libraries for each mouse included in the study and columns represent individual genes. The site at which each mouse was housed is indicated by triangles in the far left column: UT (green, left pointing) or UM (blue, right). Filled triangles correspond to those mice flagged as representative of a single B1 variant for downstream analysis. Genes are shown only where the median normalized coverage ratio between these B1-A and B1-B specific metagenomes is greater than 1.5. Rows and columns are arbitrarily ordered to maximize visual distinction between variants.

3.2.3 Genomic variation in B1

Two distinct variants of B1 were identified with one found in a majority of the UT mouse metagenomes, and the other ubiquitous at UM. Using the nucmer tool for genome alignment [35], 19.6% of B1-A MAG sequence and 12.2% of B1-B were found to not align to the other. While these hundreds of kbp may in part reflect errors in genome recovery, much of the unaligned length suggests differences in gene content between distinct sub-populations of B1. This observation was confirmed by assessing the mapping of metagenomic reads against predicted protein coding genes in each variant. For each pairing of metagenomic read library to genomic variant, gene coverage was normalized by the median gene coverage in order to identify genes with conspicuously fewer reads in particular subsets of the mice. Libraries have low coverage of large portions of either the B1-A or B1-B MAG (see Figure 3.3), suggesting that mice are primarily inhabited by one of the two variants, and that a portion of genes are variant specific.

Metagenomic libraries manually chosen as unambiguous representatives of a single B1 MAG were used to systematically identify genes differentiating the two. The median normalized mapping depths in each set of libraries against predicted genes in each MAG were compared, providing a measure of the relative enrichment or depletion of genomic sequences between the two populations of B1. This analysis found 12.8% of predicted genes in B1-A were depleted at least 5-fold in B1-B populations, and 12.4% the reverse. While this observed depletion could indicate variation in copy number, differential gene content between variants is a more parsimonious explanation for most loci. These predicted genes reflect 2.7% of unique KOs in B1-A and 1.9% in B1-B. Interestingly, the fraction of variant specific OPFs is greater, 7.5% and 7.1% respectively, suggesting that *de novo* clustering could be more sensitive to potential differences in physiology.

Table 3.2: Summary of variant specific features in two B1 MAGs

	B1-A		B1-B	
	Total	Specific	Total	Specific
Nucleotide length[1]	3.23	0.63	2.96	0.36
Genes	2,710	348	2,496	309
OPFs[2]	2,308	173	2,202	157
KOs[2]	1,056	29	1,033	20
COGs[2]	716	8	709	3

[1] in Mbp
[2] unique

Given the observation that the relative abundance of B1 was dramatically increased with ACA treatment at UM, while not being significantly affected at UT, and that B1-B was not found in metagenomes at UM, we searched for differences in functional potential between the two variants that could explain this pattern.

Genomic regions apparently specific to B1-A—defined as an at least 5-fold enrichment—include just one PUL (SusC-homolog encoded by B1A_00048). This locus includes a predicted outer membrane localized GH30 containing protein. Characterized GH30 containing proteins have β-glucosylceramidase, β-1,6-glucanase, or β-xylosidase activity [36]. Given that this PUL also encodes a periplasmic, GH3 containing protein, it appears to be unlikely that it has specificity for starch. The B1-A MAG also possesses numerous phage insertions not seen in the B1-B recon-

struction. Conversely, a CRISPR operon including 25 repeat units (Cas9 encoded by B1B_01367) appears to be specific to B1-B.

Most strikingly, a 16 kbp region (from B1A_01498 to B1A_01514) specific to B1-A was found to contain several genes with homology to cell capsule and exopolysaccharide synthesizing enzymes. Based on annotations with KEGG orthologous groups, these include homologs of *tuaG* (K16698), *tagE* (K00712), *gmhB* (K03273), *gmhA/lpcA* (K03271), *hddA* (K07031), *exoO* (K16555), *waaH* (K19354), and *tagF* (K09809). Interestingly, the B1-B MAG contains a different such region of about 6.5 kbp (B1B_00851 to B1B_00856) with *wfeD* (K21364), *pglJ* (K17248), and *epsH* (K19425). For each, several of the OPFs in the respective regions were not found anywhere in the opposing genome, suggesting that the makeup of each variant's exterior surface might be distinctly different.

3.3 Discussion

Mice are a key model system for study of the mammalian gut microbiome, with an outsized importance in testing mechanistic hypotheses for the role of this community on host health [37]. The generalizability of observations made in mice is a constant concern [37], in part due to extensive difference in taxonomic composition compared to humans [15]. The members of the *Muribaculaceae* are abundant in the murine gut microbiome [16]. While these bacteria are also found in humans (although at lower abundance), only one cultivated member of this clade has been described [15]. As a result, the ecological roles of these taxa have not been characterized, and observations in mouse model systems are therefore less valuable for understanding related processes in the human gut microbiome. Attempts to study these organisms leverage genomes reconstructed from metagenomic reads, and have proposed—in the absence of experimental data—that members of the family consume a diversity of polysaccharides in the lower gut.

Here we have extended that approach to eight new genomes, and associated those with taxa for which changes in relative abundance in response to ACA treatment have been experimentally assessed. This enabled us to explore why responders B1 and B2 each increase with ACA treatment, while the other *Muribaculaceae* do not. Annotations of reconstructed genomes suggest that these may possess starch degradation capabilities absent in the non-responders.

We examine the three-guild model proposed by Ormerod *et al.* [16] by reproducing their dimensional reduction approach with the addition of these new genomes.

In this analysis, B1 and B2 annotations appear to be consistent with a hypothesized α-glucan degradation guild, supporting their interpretation. A more nuanced approach to annotation was also applied by constructing *de novo* clusters of proteins based on homology. Interestingly, this analysis indicates that B2, and the closely related M6, share physiological potential with taxa in the host-glycan guild, suggesting that a more detailed examination can identify specific functions that discriminate responders from non-responders. This approach is bolstered by the phylogenetic and genomic distinction between B1 and B2, reducing the confounding effects of shared evolutionary history.

By including otherwise unannotated genes, genomic comparisons based on OPFs instead of previously defined gene orthologies may better reflect shared functional potential. Besides the identification of potentially novel gene families, *de novo* homology clustering [28] also enables differentiation of sub-groups not captured by standard annotations. For instance, hypothetical genes annotated as homologs of SusC, SusD, and SusEF, were clustered into 119, 162, and 33 different OPFs respectively. It is plausible that this sub-clustering captures differences in protein structure with importance in oligo- and polysaccharide recognition, import, and binding. Combined with annotation of characterized functional domains, these clusters may better predict the polysaccharide utilization ranges of uncultured organisms.

A detailed analysis of PULs identified multiple loci in B1 that appear to be adapted to the degradation of starch or related carbohydrates, due to the presence of an OM localized GH13 containing protein [38]. Counterintuitively, B2 had no such PUL, suggesting that its response to ACA may result from other enzymatic capabilities. Of particular interest is a PUL encoding proteins with GH97, CBM20, and CBM69 domains, all of which have documented activity on starch [39, 40]. While the only outer-membrane localized hydrolase in this PUL is a GH66, and members of this family have characterized activity on the α-1,6 linkages between glucose monomers in dextran [41]. It is plausible that this PUL can be repurposed and confers some ability to grow on starch.

In addition, a gene encoding a SusA homolog was identified in both B1 and B2 but in none of the non-responders. While it is unclear how expression of this important component of starch utilization might be regulated, given that it is not located in a PUL in either of the responders, SusA is important for growth on amylopectin in *B. thetaiotaomicron* [34]. Since inhibition by acarbose is variable across enzymes [42], it is possible that acarbose treatment results in elevated levels of dextrin and maltooligosaccharides in the lower guts of mice due to residual α-amylase activity, even

at levels sufficient to prohibit host digestion. Periplasmic hydrolysis of these starch breakdown products may be sufficient for increased abundance of these taxa in acarbose treated mice.

It is notable that two distinct variants of B1 were identifiable in these metagenomes, and that the distribution of B1-A and B1-B are reminiscent of the previously observed site-specificity of ACA response. Despite evidence that genomic variation is common in the bacterial world [43, 44], studies reconstructing genomes from metagenomes often ignore this possibility (with a few notably exceptions [45, 46]). The discovery of two subpopulations of B1 therefore demonstrates the value of considering pangenome dynamics, and presents a potential explanation for the observed site-specific response of that taxon. The finding that both variants have the same complement of three PULs apparently specializing in starch utilization and the same SusA homolog does not support the hypothesis that differences in starch utilization potential account for these abundance patterns. We did, however, identify numerous differences in the gene content of B1-A and B1-B, including variant specific loci that may influence the structure and function of the outer surface of the cell. Capsule variation is known to greatly affect both ecological and host interactions [47].

While these results do not establish a mechanistic explanation for differences in the response of B1 at UM and UT, nor conclusively identify starch utilization pathways in B2, they do suggest a number of genomic features that likely contribute to previously observed patterns in taxon abundance. Future studies utilizing metatranscriptomic analysis might demonstrate active expression of these genes, or differential expression in mice treated with acarbose compared to controls. Likewise, even in the absence of a B2 cultivar, the sufficiency of the dextran PUL for increased growth with acarbose treatment could be tested using available cultivars, including *B. thetaiotaomicron*.

3.4 Conclusions

In this study we have reconstructed and described genomes representing 7 OTUs in the family *Muribaculaceae* from the mouse fecal microbiome, and have found features that differentiate those that respond positively to ACA treatment from those that do not. This analysis suggests that utilization of starch and related polysaccharides enables increased population size in mice treated with the α-amylase inhibitor. In addition, two distinct genomic variants of one taxon were identified that differ in functional gene content, potentially explaining site-specific differences in response. By combining observed changes in relative abundance during experimental manipulation

with inferred functional gene content, we are able to study mammalian symbionts in the absence of cultured representatives. This sequence-based approach is broadly applicable in microbial ecology and enables improved understanding of *in situ* dynamics within complex microbial communities.

3.5 Methods

3.5.1 Mouse treatment, sample collection, extraction and sequencing

Mice were bred, housed, and treated as described in [13]. Briefly, genetically heterogeneous UM-HET3 mice at each study site were produced by the four-way cross detailed in [48]. Mice were fed LabDiet (TestDiet Inc.) 5LG6 from weaning onwards. Starting at 8 months of age, mice randomly assigned to treatment were fed chow with 1,000 ppm ACA (Spectrum Chemical Manufacturing Corporation). Mice were housed 4 males or 5 females to a cage. Colonies were assessed for infectious agents every 3 months, and all tests were negative.

Individual fecal pellets were collected from a single mouse per cage. 16S rRNA gene libraries and metabolite analyses of these samples are described in Chapter 2. From this collection, a subset of samples were non-randomly selected for metagenomic sequencing based on various criteria. Samples were from 54 mice, with at least six treated and control representatives of both males and females at each site.

Fecal samples were slurried with nuclease free water at a 1:10 (w/v) ratio, and most samples were spiked with *Sphingopyxis alaskensis* RB2256 prepared as described in Chapter 2 before DNA extraction and sequencing. Based on alignment to the reference genome, sequenced reads from *S. alaskensis* can be distinguished from all endogenous bacteria in mouse feces. A small number of these were split for both spiked and unspiked samples, which we used to validate this procedure. For each, 150 μLof this sample was transferred for extraction using the MoBio PowerMag Microbiome kit. Metagenomic libraries were prepared using standard procedures sequenced on the Illumina HiSeq 400 platform using the v4 paired-end 2x150 bp.

3.5.2 Assembly, binning, and MAG refinement

Raw metagenomic reads were deduplicated using FastUniq [49], adapters trimmed using Scythe [50], and quality trimmed using Sickle [51] to produce processed reads for

all downstream analyses. The resulting paired-end reads were assembled into primary contigs using MEGAHIT [52]. Reads were then mapped back to these contigs with Bowtie2 [53], and per-library coverage was estimated for each contig.

For all contigs >1000 bp in length, dimensional reductions built into CONCOCT [54] were applied to produce input data for a Gaussian mixture model (GMM) similar to the procedure used by that program for binning. However, unlike CONCOCT—due to computational limitations—the model was trained on just 10% of the input data, sampled randomly, before assigning bins to all contig. While this may have reduced the accuracy of the binning procedure, we believe that subsequent refinement steps mitigated the impact of this decision.

OTUs were classified taxonomically and relative abundance was calculated for matched libraries as described in Chapter 2. Bins were then recruited to one or more OTUs by calculating a Canonical partial least squares between OTU abundance and bin coverage as implemented in the scikit-learn machine learning library for Python [55]. For bins recruited to OTUs classified as *Muribaculaceae*, contigs were re-clustered based on coverage across samples. First "trusted contigs" were manually selected which correlated closely with OTU abundance. The mean coverage of these was used to normalize the per-library coverage of all other contigs. Then, using a GMM, groups of contigs were clustered such that the normalized coverage across samples was consistent. These groups were used to inform the manual assignment of contigs to MAGs. Libraries in which MAGs had non-negligible coverage were identified and used in subsequent refinements. For the B1 reconstruction, but no other MAGs, a number of groups containing on the order of 10^5 bp were found with low coverage in just a subset of libraries. By this criterion, contigs in these "variable" groups were partitioned into two MAG variants, A and B, with non-variable groups shared by both. Only libraries that appeared on further inspection to have just one of the two variants were considered in downstream refinement steps. The mice matching these libraries are highlighted in Figure 3.3.

For each MAG, several alternative refinement procedures were performed from which the best quality result was selected. Reads mapping to the curated contigs were digitally normalized [56–58] and reassembled with SPADES [59]. This reassembly as well as the original contigs were cleaned using a single pass of the Pilon assembly refinement tool [60]. Finally, the per-library mapping depths of each position in these assemblies were compared to the average mapping depth of the "trusted contigs" selected earlier, and regions with low cosine similarity were excised from the final assemblies.

Genome completeness and contamination estimates were calculated based on ubiquitous single-copy genes using the program CheckM [26]. Based on these results, the final assembly with the highest completeness and with contamination $< 1\%$ was selected from the various refinements.

3.5.3 Reference genomes

The *Muribaculum intestinale* genome sequence was obtained from GenBank (accession GCA_002201515.1), as well as four additional draft genomes (GCA_003024805.1, GCA_003024815.1, GCA_002633305.1, GCA_002633115.1). While other genomes labeled as *Muribaculaceae* have also been deposited, they were excluded from this analysis due to redundancy or apparent misidentification to the family. The 30 MAGs reconstructed by Ormerod *et al.* [16] were obtained from the SRA. For comparison, nucleotide sequences for *B. thetaiotaomicron* VPI-5482 (AE015928.1), *B. ovatus* (CP012938.1), *Barnesiella viscericola* (GCA_000512915.1), and *Porphyromonas gingivalis* (GCA_000010505.1), were also downloaded from GenBank.

3.5.4 Genome annotation

All genomes were initially annotated with Prokka [61] version 1.13, which uses Prodigal [62] for gene finding. Putative protein sequences were additionally annotated with domains from both the dbCAN database [23] release 6 of carbohydrate-active domains and Pfam [63] release 31.0, using HMMER3 [64, 65] version 3.1b2. Protein sequences were also annotated with KO numbers by BLAST using the KEGG database as of March 2018 as the reference and taking the best hit with a maximum E-value of 1e-10.

Lipoproteins were predicted using LipoP [66] (version 1.0a) and a score cutoff of 5 and a margin cutoff of 2. Lipoproteins with an arginine at position +2 relative to the cleavage site were labeled as localized to the inner membrane. Periplasmic proteins were identified with SignalP [67] (version 4.1). Predicted protein sequences from all annotated genomes were locally all-by-all aligned using the DIAMOND implementation of the BLAST algorithm [68]. Each pair was then assigned a similarity value as the bitscore of their best local alignment normalized by the greater of the two self-alignments. This results in a matrix of pairwise scores reflecting the proximity to perfect homology. Scores less than 0.2 were replaced with 0. Clusters were formed using the MCL algorithm [69] with an inflation parameter of 5.

SusCDEF homologs were identified based on relatively relaxed criteria, harnessing OPF assignments, domain predictions, and Prokka annotations to avoid false nega-

tives while maintaining specificity. For each OPF, all KOs assigned to members were collected as plausible KOs for the cluster. Protein sequences in OPF clusters which included K21572 were flagged as putative SusC-homologs, as were sequences directly annotated as such by Prokka. Using a similar approach, proteins in clusters tagged with K21571 or with any of domains PF12771, PF14322, PF12741, PF07980 were identified as putative SusD. Proteins in clusters tagged with K21571, or with either PF14292 or PF16411, were considered SusEF homologs. PULs were identified by a SusC-homolog with its start codon within 5 kbp of a SusD-homolog's start on the same strand. Manual inspection supported the vast majority of these identifications.

BIBLIOGRAPHY

[1] Turnbaugh, P.J., Ley, R.E., Mahowald, M.A.: An obesity-associated gut microbiome with increased capacity for energy harvest. Nature **444** (2006)

[2] Britton, R.A., Young, V.B.: Interaction between the intestinal microbiota and host in Clostridium difficile colonization resistance. Trends in Microbiology **20**(7), 313–9 (2012). doi:10.1016/j.tim.2012.04.001

[3] Syal, G., Kashani, A., Shih, D.Q.: Fecal Microbiota Transplantation in Inflammatory Bowel Disease- a Primer for the Internists. The American Journal of Medicine (2018). doi:10.1016/j.amjmed.2018.03.010

[4] Hiele, M., Ghoos, Y., Rutgeerts, P., Vantrappen, G.: Effects of acarbose on starch hydrolysis. Digestive Diseases and Sciences **37**(7), 1057–1064 (1992). doi:10.1007/BF01300287

[5] Dehghan-Kooshkghazi, M., Mathers, J.C.: Starch digestion, large-bowel fermentation and intestinal mucosal cell proliferation in rats treated with the α-glucosidase inhibitor acarbose. British Journal of Nutrition **91**(03), 357 (2004). doi:10.1079/BJN20031063

[6] Zhao, L., Zhang, F., Ding, X., Wu, G., Lam, Y.Y., Shi, Y., Shen, Q., Dong, W., Liu, R., Ling, Y., Zeng, Y.: Gut bacteria selectively promoted by dietary fibers alleviate type 2 diabetes. Science **1156**(March), 1151–1156 (2018). doi:10.1126/science.aao5774

[7] Holt, P.R., Atillasoy, E., Lindenbaum, J., Ho, S.B., Lupton, J.R., McMahon, D., Moss, S.F.: Effects of acarbose on fecal nutrients, colonic pH, and short-chain fatty acids and rectal proliferative indices. Metabolism: Clinical and Experimental **45**(9), 1179–1187 (1996). doi:10.1016/S0026-0495(96)90020-7

[8] Wolever, T.M.S., Chiasson, J.L.: Acarbose raises serum butyrate in human subjects with impaired glucose tolerance. British Journal of Nutrition **84**(1), 57–61 (2000). doi:10.1017/S0007114500001239

[9] Weaver, G.A., Tangel, C.T., Krause, J.A., Parfitt, M.M., Jenkins, P.L., Rader, J.M., Lewis, B.A., Miller, T.L., Wolin, M.J.: Acarbose Enhances Human Colonic Butyrate Production. The Journal of Nutrition **127**(5), 717–723 (1997). doi:10.1093/jn/127.5.717

[10] Weaver, G.A., Tangel, C.T., Krause, J.A., Parfitt, M.M., Stragand, J.J., Jenkins, P.L., Erb, T.A., Davidson, R.H., Alpern, H.D., Guiney, W.B., Higgins, P.J.: Biomarkers of human colonic cell growth are influenced differently by a history of colonic neoplasia and the consumption of acarbose. The Journal of Nutrition **130**(11), 2718–25 (2000). doi:10.1093/jn/130.11.2718

[11] Wolin, M.J., Miller, T.L., Yerry, S., Bank, S., Weaver, G.A., Zhang, Y.: Changes of Fermentation Pathways of Fecal Microbial Communities Associated with a Drug Treatment That Increases Dietary Starch in the Human Colon Changes of Fermentation Pathways of Fecal Microbial Communities Associated with a Drug Treatment That Increas. Applied and Environmental Microbiology **65**(7), 2807–2812 (1999)

[12] Zhang, X., Fang, Z., Zhang, C., Xia, H., Jie, Z., Han, X., Chen, Y., Ji, L.: Effects of Acarbose on the Gut Microbiota of Prediabetic Patients: A Randomized, Double-blind, Controlled Crossover Trial. Diabetes Therapy **8**(2), 293–307 (2017). doi:10.1007/s13300-017-0226-y

[13] Harrison, D.E., Strong, R., Allison, D.B., Ames, B.N., Astle, C.M., Atamna, H., Fernandez, E., Flurkey, K., Javors, M.A., Nadon, N.L., Nelson, J.F., Pletcher, S., Simpkins, J.W., Smith, D.L., Wilkinson, J.E., Miller, R.A.: Acarbose, 17-α-estradiol, and nordihydroguaiaretic acid extend mouse lifespan preferentially in males. Aging Cell **13**(2), 273–282 (2014). doi:10.1111/acel.12170

[14] Strong, R., Miller, R.A., Antebi, A., Astle, C.M., Bogue, M., Denzel, M.S., Fernandez, E., Flurkey, K., Hamilton, K.L., Lamming, D.W., Javors, M.A., de Magalhães, J.P., Martinez, P.A., McCord, J.M., Miller, B.F., Müller, M., Nelson, J.F., Ndukum, J., Rainger, G.E., Richardson, A., Sabatini, D.M., Salmon, A.B., Simpkins, J.W., Steegenga, W.T., Nadon, N.L., Harrison, D.E.: Longer lifespan in male mice treated with a weakly estrogenic agonist, an antioxidant, an α-glucosidase inhibitor or a Nrf2-inducer. Aging Cell **15**(5), 872–884 (2016). doi:10.1111/acel.12496

[15] Lagkouvardos, I., Pukall, R., Abt, B., Foesel, B.U., Meier-Kolthoff, J.P., Kumar, N., Bresciani, A., Martínez, I., Just, S., Ziegler, C., Brugiroux, S., Garzetti, D., Wenning, M., Bui, T.P.N., Wang, J., Hugenholtz, F., Plugge, C.M., Peterson, D.A., Hornef, M.W., Baines, J.F., Smidt, H., Walter, J., Kristiansen, K., Nielsen, H.B., Haller, D., Overmann, J., Stecher, B., Clavel, T.: The Mouse Intestinal

Bacterial Collection (miBC) provides host-specific insight into cultured diversity and functional potential of the gut microbiota. Nature Microbiology **1**(August), 16131 (2016). doi:10.1038/nmicrobiol.2016.131

[16] Ormerod, K.L., Wood, D.L.A., Lachner, N., Gellatly, S.L., Daly, J.N., Parsons, J.D., Dal'Molin, C.G.O., Palfreyman, R.W., Nielsen, L.K., Cooper, M.A., Morrison, M., Hansbro, P.M., Hugenholtz, P.: Genomic characterization of the uncultured Bacteroidales family S24-7 inhabiting the guts of homeothermic animals. Microbiome **4**(1), 36 (2016). doi:10.1186/s40168-016-0181-2

[17] Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.A., Woodcroft, B.J., Evans, P.N., Hugenholtz, P., Tyson, G.W.: Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nature Microbiology **2**(11), 1533–1542 (2017). doi:10.1038/s41564-017-0012-7

[18] Lee, S.T.M., Kahn, S.A., Delmont, T.O., Shaiber, A., Esen, ö.C., Hubert, N.A., Morrison, H.G., Antonopoulos, D.A., Rubin, D.T., Eren, A.M.: Tracking microbial colonization in fecal microbiota transplantation experiments via genome-resolved metagenomics. Microbiome **5**(1) (2017). doi:10.1186/S40168-017-0270-X. arXiv:1011.1669v3

[19] Martens, E.C., Koropatkin, N.M., Smith, T.J., Gordon, J.I.: Complex glycan catabolism by the human gut microbiota: The bacteroidetes sus-like paradigm. Journal of Biological Chemistry **284**(37), 24673–24677 (2009). doi:10.1074/jbc.R109.022848

[20] Foley, M.H., Cockburn, D.W., Koropatkin, N.M.: The Sus operon: a model system for starch uptake by the human gut Bacteroidetes. Cellular and Molecular Life Sciences **73**(14), 2603–2617 (2016). doi:10.1007/s00018-016-2242-x

[21] Grondin, J.M., Tamura, K., Déjean, G., Abbott, D.W., Brumer, H.: Polysaccharide utilization loci: Fueling microbial communities. Journal of Bacteriology **199**(15) (2017). doi:10.1128/JB.00860-16

[22] Fernández-Gómez, B., Richter, M., Schüler, M., Pinhassi, J., Acinas, S.G., González, J.M., Pedrós-Alió, C.: Ecology of marine bacteroidetes: A comparative genomics approach. ISME Journal **7**(5), 1026–1037 (2013). doi:10.1038/ismej.2012.169

[23] Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F., Xu, Y.: DbCAN: A web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Research **40**(W1), 445–451 (2012). doi:10.1093/nar/gks479

[24] Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z., Busk, P.K., Xu, Y., Yin, Y.: DbCAN2: A meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Research **46**(W1), 95–101 (2018). doi:10.1093/nar/gky418

[25] Stewart, E.J.: Growing unculturable bacteria. Journal of Bacteriology **194**(16), 4151–4160 (2012). doi:10.1128/JB.00345-12. arXiv:1011.1669v3

[26] Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., Tyson, G.W.: CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Research **25**(7), 1043–1055 (2015). doi:10.1101/gr.186072.114

[27] Kim, B.J., Hong, S.H., Kook, Y.H., Kim, B.J.: Molecular Evidence of Lateral Gene Transfer in rpoB Gene of Mycobacterium yongonense Strains via Multilocus Sequence Analysis. PLoS ONE **8**(1), 1–5 (2013). doi:10.1371/journal.pone.0051846

[28] Schloss, P.D., Handelsman, J.: A statistical toolbox for metagenomics: Assessing functional diversity in microbial communities. BMC Bioinformatics **9**, 1–15 (2008). doi:10.1186/1471-2105-9-34

[29] Shipman, J.A., Cho, K.H., Siegel, H.A., Salyers, A.A.: Physiological characterization of SusG, an outer membrane protein essential for starch utilization by Bacteroides thetaiotaomicron. Journal of Bacteriology **181**(23), 7206–7211 (1999)

[30] Janeček, Š., Svensson, B., MacGregor, E.A.: α-Amylase: an enzyme specificity found in various families of glycoside hydrolases. Cellular and Molecular Life Sciences **71**, 1149–1170 (2014). doi:10.1007/s00018-013-1388-z

[31] Ravcheev, D.A., Godzik, A., Osterman, A.L., Rodionov, D.A.: Polysaccharides utilization in human gut bacterium Bacteroides thetaiotaomicron: comparative genomics reconstruction of metabolic and regulatory networks. BMC Genomics **14**(1), 873 (2013). doi:10.1186/1471-2164-14-873

[32] Rogers, T.E., Pudlo, N.A., Koropatkin, N.M., Bell, J.S.K., Moya Balasch, M., Jasker, K., Martens, E.C.: Dynamic responses of Bacteroides thetaiotaomicron during growth on glycan mixtures. Molecular Microbiology **88**(5), 876–890 (2013). doi:10.1111/mmi.12228. NIHMS150003

[33] van Bueren, A.L., Saraf, A., Martens, E.C., Dijkhuizena, L.: Differential metabolism of exopolysaccharides from probiotic lactobacilli by the human gut symbiont Bacteroides thetaiotaomicron. Applied and Environmental Microbiology **81**(12), 3973–3983 (2015). doi:10.1128/AEM.00149-15

[34] D'Elia, J.N., Salyers, A.A.: Contribution of a neopullulanase, a pullulanase, and an ??-glucosidase to growth of Bacteroides thetaiotaomicron on starch. Journal of Bacteriology **178**(24), 7173–7179 (1996). doi:10.1016/j.tim.2004.07.004

[35] Delcher, A.L.: Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Research **30**(11), 2478–2483 (2002). doi:10.1093/nar/30.11.2478

[36] St John, F.J., González, J.M., Pozharski, E.: Consolidation of glycosyl hydrolase family 30: A dual domain 4/7 hydrolase family consisting of two structurally distinct groups. FEBS Letters **584**(21), 4435–4441 (2010). doi:10.1016/j.febslet.2010.09.051

[37] Nguyen, T.L.A., Vieira-Silva, S., Liston, A., Raes, J.: How informative is the mouse for human gut microbiota research? Disease Models & Mechanisms **8**(1), 1–16 (2015). doi:10.1242/dmm.017400

[38] Koropatkin, N.M., Smith, T.J.: SusG: A Unique Cell-Membrane-Associated α-Amylase from a Prominent Human Gut Symbiont Targets Complex Starch Molecules. Structure **18**(2), 200–215 (2010). doi:10.1016/j.str.2009.12.010

[39] Naumoff, D.G.: GH97 is a new family of glycoside hydrolases, which is related to the α-galactosidase superfamily. BMC Genomics **6**, 1–12 (2005). doi:10.1186/1471-2164-6-112

[40] Boraston, A.B., Bolam, D.N., Gilbert, H.J., Davies, G.J.: Carbohydrate-binding modules: fine-tuning polysaccharide recognition. Biochemical Journal **382**(3), 769–781 (2004). doi:10.1042/BJ20040892. arXiv:1011.1669v3

[41] Kim, Y.M., Yamamoto, E., Kang, M.S., Nakai, H., Saburi, W., Okuyama, M., Mori, H., Funane, K., Momma, M., Fujimoto, Z., Kobayashi, M., Kim, D., Kimura, A.: Bacteroides thetaiotaomicron VPI-5482 glycoside hydrolase family 66 homolog catalyzes dextranolytic and cyclization reactions. FEBS Journal **279**(17), 3185–3191 (2012). doi:10.1111/j.1742-4658.2012.08698.x

[42] Kim, M.J., Lee, S.B., Lee, H.S., Lee, S.Y., Baek, J.S., Kim, D., Moon, T.W., Robyt, J.F., Park, K.H.: Comparative study of the inhibition of alpha-glucosidase, alpha-amylase, and cyclomaltodextrin glucanosyltransferase by acarbose, isoacarbose, and acarviosine-glucose. Archives of biochemistry and biophysics **371**(2), 277–283 (1999). doi:10.1006/abbi.1999.1423

[43] Rasko, D.A., Rosovitz, M.J., Myers, G.S.A., Mongodin, E.F., Fricke, W.F., Gajer, P., Crabtree, J., Sebaihia, M., Thomson, N.R., Chaudhuri, R., Henderson, I.R., Sperandio, V., Ravel, J.: The pangenome structure of Escherichia coli: Comparative genomic analysis of E. coli commensal and pathogenic isolates. Journal of Bacteriology **190**(20), 6881–6893 (2008). doi:10.1128/JB.00619-08. Rasko, David A, 2008, The Pangenome

[44] Medini, D., Donati, C., Tettelin, H., Masignani, V., Rappuoli, R.: The microbial pan-genome. Current Opinion in Genetics and Development **15**(6), 589–594 (2005). doi:10.1016/j.gde.2005.09.006. Masignani, Vega, 2005, The microbial

[45] Truong, D.T., Tett, A., Pasolli, E., Huttenhower, C., Segata, N.: Microbial strain-level population structure and genetic diversity from metagenomes. Genome Research **27**(4), 626–638 (2017). doi:10.1101/gr.216242.116

[46] Delmont, T.O., Eren, A.M.: Linking pangenomes and metagenomes: the Prochlorococcus metapangenome. PeerJ **6**, 4320 (2018). doi:10.7717/peerj.4320

[47] Merino, S., Tomás, J.M.: Bacterial Capsules and Evasion of Immune Responses. eLS, 1–10 (2015). doi:10.1002/9780470015902.a0000957.pub4

[48] Miller, R.A., Harrison, D.E., Astle, C.M., Baur, J.A., Boyd, A.R., de Cabo, R., Fernandez, E., Flurkey, K., Javors, M.a., Nelson, J.F., Orihuela, C.J., Pletcher, S., Sharp, Z.D., Sinclair, D.A., Starnes, J.W., Wilkinson, J.E., Nadon, N.L., Strong, R.: Rapamycin, but not resveratrol or simvastatin, extends life span of genetically heterogeneous mice. Journals of Gerontology - Series A Biological Sciences and Medical Sciences **66 A**(2), 191–201 (2011). doi:10.1093/gerona/glq178

[49] Xu, H., Luo, X., Qian, J., Pang, X., Song, J., Qian, G., Chen, J., Chen, S.: FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads. PLoS ONE **7**(12), 1–6 (2012). doi:10.1371/journal.pone.0052249

[50] Buffalo, V.: Scythe: A 3'-end adapter contaminant trimmer (2018)

[51] Joshi, N.A., Fass, J.N.: Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) (2011). https://github.com/najoshi/sickle

[52] Li, D., Liu, C.M., Luo, R., Sadakane, K., Lam, T.W.: MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics **31**(10), 1674–1676 (2014). doi:10.1093/bioinformatics/btv033. 1401.7457

[53] Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with Bowtie 2. Nature Methods **9**(4), 357–359 (2012). doi:10.1038/nmeth.1923. #14603

[54] Alneberg, J., Bjarnason, B.S., De Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., Lahti, L., Loman, N.J., Andersson, A.F., Quince, C.: Binning metagenomic contigs by coverage and composition. Nature Methods **11**(11), 1144–1146 (2014). doi:10.1038/nmeth.3103

[55] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011). 1201.0490

[56] Wedemeyer, A., Kliemann, L., Srivastav, A., Schielke, C., Reusch, T.B., Rosenstiel, P.: An improved filtering algorithm for big read datasets and its application to single-cell assembly. BMC Bioinformatics **18**(1), 1–11 (2017). doi:10.1186/s12859-017-1724-7. 1610.03443

[57] Brown, C.T., Howe, A., Zhang, Q., Pyrkosz, A.B., Brom, T.H.: A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data, 1–18 (2012). doi:10.1128/genomeA.00802-14.Copyright. 1203.4802

[58] Zhang, Q., Pell, J., Canino-Koning, R., Howe, A.C., Brown, C.T.: These are not the K-mers you are looking for: Efficient online K-mer counting using a probabilistic data structure. PLoS ONE **9**(7) (2014). doi:10.1371/journal.pone.0101271. 1309.2975

[59] Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A.: SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology **19**(5), 455–477 (2012). doi:10.1089/cmb.2012.0021. 1604.03071

[60] Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., Earl, A.M.: Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE **9**(11) (2014). doi:10.1371/journal.pone.0112963

[61] Seemann, T.: Prokka: Rapid prokaryotic genome annotation. Bioinformatics **30**(14), 2068–2069 (2014). doi:10.1093/bioinformatics/btu153. arXiv:1401.4290v2

[62] Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J.: Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics **11**(1), 119 (2010). doi:10.1186/1471-2105-11-119. 1401.7457

[63] Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E.L.L., Eddy, S.R., Bateman, A., Finn, R.D.: The Pfam protein families database. Nucleic acids research **40**(Database issue), 290–301 (2012). doi:10.1093/nar/gkr1065

[64] Eddy, S.R.: Accelerated Profile HMM Searches. PLoS computational biology **7**(10), 1002195 (2011). doi:10.1371/journal.pcbi.1002195

[65] Eddy, S.R.: A new generation of homology search tools based on probabilistic inference. Genome informatics. International Conference on Genome Informatics **23**(1), 205–11 (2009)

[66] Juncker, A.S., Willenbrock, H., von Heijne, G., Brunak, S., Nielsen, H., Krogh, A.: Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Science **12**(8), 1652–1662 (2003). doi:10.1110/ps.0303703

[67] Petersen, T.N., Brunak, S., Von Heijne, G., Nielsen, H.: SignalP 4.0: Discriminating signal peptides from transmembrane regions. Nature Methods **8**(10), 785–786 (2011). doi:10.1038/nmeth.1701

[68] Buchfink, B., Xie, C., Huson, D.H.: Fast and sensitive protein alignment using DIAMOND. Nature Methods **12**(1), 59–60 (2014). doi:10.1038/nmeth.3176

[69] Enright, A.J., Van Dongen, S., Ouzounis, C.A.: An efficient algorithm for large-scale detection of protein families. Nucleic acids research **30**(7), 1575–1584 (2002). doi:10.1093/nar/30.7.1575. journal.pone.0035671

CHAPTER 4

Experimental considerations for spike-in quantification of absolute abundance in microbial ecology

4.1 Background

The central role of bacterial communities in processes as disparate and important as global geochemistry, waste-water treatment, and human digestion is undisputed. Contemporary approaches to understanding these often complex and cryptic ecosystems rely heavily on culture-independent, sequence-based surveys of taxonomic marker genes, most notably the 16S rRNA gene [1]. In studies of human health, these methods have yielded discoveries of associations between community composition and numerous biological outcomes including colorectal cancer [2, 3], obesity [4], psoriasis [5], and autism [6]. Despite this progress, demonstrating mechanistic roles for bacteria in health and disease remains a major challenge.

One barrier to moving beyond simple associations is the disconnect between measurements of community makeup obtained from community surveys and the underlying microbial population size. Due to variation in quantities of biological sample, variation in DNA extraction and amplification efficiency, gene copy number variation, and library normalization, the number of reads recovered from a particular taxon is only meaningful when normalized to library size and treated as a proxy for relative abundance [7]. Such measures always sum to 100%, a feature of this type of data termed "compositionality" [8]. However, metabolic rates, toxin production, and microbial biomass, for instance, should be expected to scale with absolute abundance rather than relative abundance of the relevant taxa [e.g. 9, 10]. When total community size is not constant, changes in composition do not necessarily reflect equivalent changes in cell counts or biomass.

To escape this limitation, direct cell counts, biomolecule quantification, and qPCR, have been used to transform relative abundance data to a proxy for absolute abundance [methods compared in 11] However, each of these methods requires additional processing for every sample, which can be expensive, time intensive, and consume valuable sample material, explaining the rarity of these approaches in published studies.

Recently, an alternative approach has been proposed, here referred to as spike-in quantification [12, 13]. By adding known amounts of a recognizable DNA sequence to samples before extraction, the total size of the endogenous community may be estimated based on the recovery of this foreign sequence. While demonstrations of this method have been published, standard protocols for sample handling, spiking, and bioinformatic processing have not been proposed.

Here we discuss spike-in quantification and argue for its broader application in microbial community analysis. This chapter introduces a conceptual framework for the approach motivating its application and forming a basis for future developments. Towards this end, a brief review of relevant literature is included, along with several new experimental results, as well as observations from the author's experience with its application. Section 4.2 explains the shortcomings of compositional data as well as current methods for the estimation of absolute abundance. Section 4.3 provides a conceptual overview of spike-in quantification, describes the risks, and shares new experimental data demonstrating the robustness of the approach to several plausible sources of technical variability. Section 4.4 presents a comprehensive set of suggestions for spike-in quantification protocols. And Section 4.5 describes remaining challenges not solved by spike-in quantification.

By leveraging spike-in quantification, sequence-based microbial community surveys can be closer aligned with biological reality, enabling mechanistic insights inaccessible to current techniques.

4.2 Limitations of existing approaches

Contemporary approaches to microbial community analysis are largely grounded in high-throughput sequencing of whole-community 16S rRNA gene libraries. The data generated by these methods are analyzed and interpreted in units of percent or fractional abundance of individual taxonomic marker genes, a proxy for the relative abundance of organisms in the initial sample. Such measures of relative abundance always sum to 100%. For this reason, increases in the abundance of one taxon are necessarily

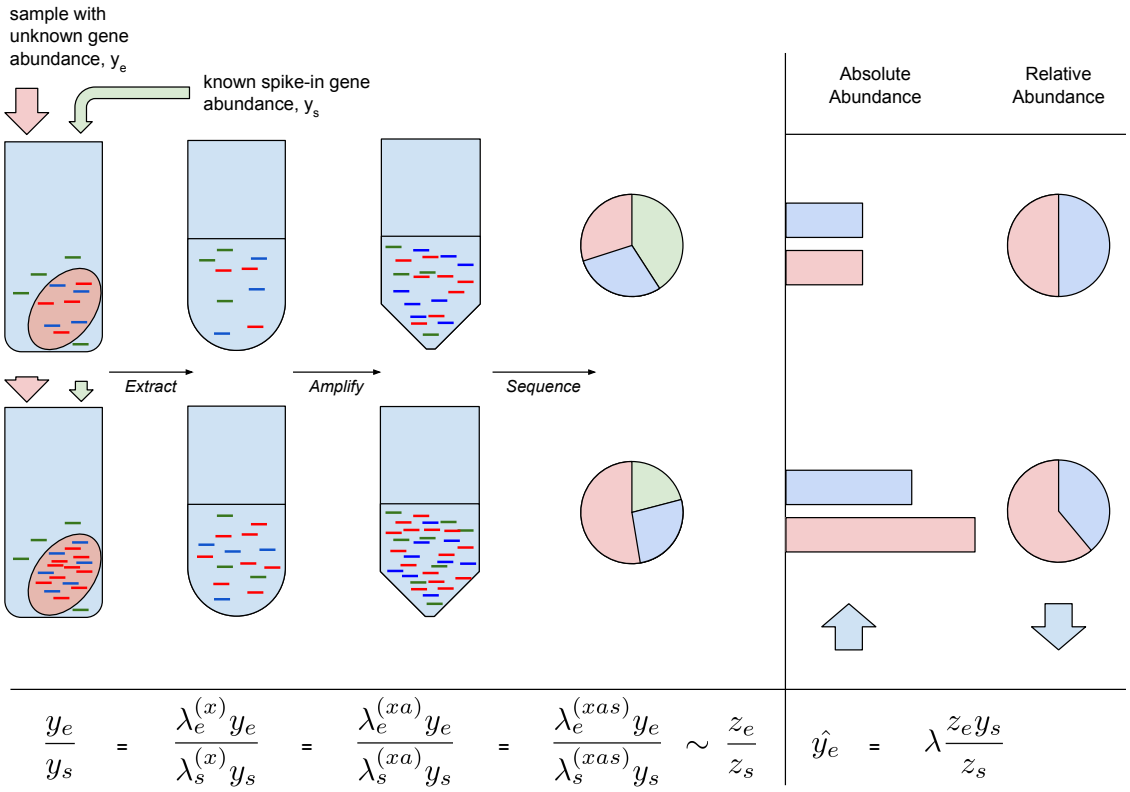
sample with unknown gene abundance, y_e
known spike-in gene abundance, y_s
Extract
Amplify
Sequence
Absolute Abundance
Relative Abundance
$\frac{y_e}{y_s} = \frac{\lambda_e^{(x)} y_e}{\lambda_s^{(x)} y_s} = \frac{\lambda_e^{(xa)} y_e}{\lambda_s^{(xa)} y_s} = \frac{\lambda_e^{(xas)} y_e}{\lambda_s^{(xas)} y_s} \sim \frac{z_e}{z_s}$
$\hat{y}_e = \lambda \frac{z_e y_s}{z_s}$

Figure 4.1: Conceptual overview of spike-in quantification. Protocols begin with the addition of a spike reagent containing a known abundance of foreign genes (green lines) to a sample containing an unknown abundance of endogenous genes (red and blue lines). Both experience the same extraction, amplification, and sequencing steps. After sequencing, a smaller relative abundance of spike reads corresponds with a larger endogenous community size. Changes in total abundance can limit inferences drawn from relative abundance measurements. When comparing the abundance of genes for the blue taxon in a sample with increased total community size (bottom series) to a reference (top series), a decreased relative abundance may counterintuitively coincide with a greater actual density of that taxon. Mathematical expressions describe the ratio between the abundance of endogenous (y_e) and spike genes (y_s) at each step, based on the various λ coefficients, which describe the cumulative extraction ($\lambda^{(x)}$), amplification ($\lambda^{(xa)}$), and sequencing efficiency ($\lambda^{(xas)}$) of the spike and endogenous genes, independently. These are reduced to a single calibration coefficient, λ, in the final formula. The absolute abundance of endogenous bacteria can be estimated from the read counts of endogenous (z_e) and spike (z_s) genes.

linked with decreased relative abundance of one or more others, and changes in relative abundance should not, on their own, be interpreted as changes in the population density of that organism.

While, microbial community analysis has adapted to the compositionality of amplicon survey data by adopting the appropriate statistical tools and carefully limiting interpretations [14, 15], it is unclear whether relative abundance is generally relevant to the biological processes carried out by bacteria. Instead, metabolic rates likely scale with the actual density of organisms in the sample, and not their relative abundance. Inference is particularly handicapped in scenarios with systematic differences in total community size, for instance studies on the effect of antibiotic treatment [16] or probiotic supplements [17]. Additionally, the goal of inferring ecological interactions from community survey data is hindered by implicit negative correlations in relative abundance [18]. Available approaches either ignore this systematic problem [19] or depend on prior belief about the sparsity of interaction networks [20–22] to circumvent the limitation.

Changes in relative abundance should only be interpreted as changes in population when additional information about total community size is available. Total bacterial density can be directly estimated using a variety of methods including direct cell counts[23–26], qPCR for ubiquitous genes [27, 28], or quantification of biomolecules [11, 29–32]. These approaches have been particularly popular in biogeochemistry, where bacterial biomass and metabolic activity are often inputs to predictive models [e.g. 10, 33, 34]. In studies of host-associated microbial communities, however, where metabolic models are less common, quantification methods have only been applied in a limited number of studies [24]. This may be due to the significant additional time and expense required.

Compared to most other options, qPCR is a relatively simple solution as it can be performed on the same DNA extracted for community sequencing. However, such quantification of community size depends crucially on the constant efficiency of extraction itself. Extraction efficiency varies between protocols [35] and commercially available DNA extraction kits commonly used for community analysis are not designed for quantitative extraction; i.e. the concentration of DNA extracted is not necessarily proportional to the quantity of DNA in the initial sample [27]. Differences in lysis across taxa is a well accepted feature of community extraction protocols [36]. Additionally, PCR amplification is inhibited by a variety of compounds that are not removed during DNA extraction and may be found in variable concentrations across samples [37]. As an added challenge, measuring the mass or volume of the small

sample aliquots that DNA is extracted from is not trivial, adding measurement error when normalizing absolute abundance to the sample quantity in order to get a comparable measure of population density. These technical limitations do not detract from the use of marker gene surveys as a proxy for relative abundance as long as differences in extraction efficiency between samples affect all taxa equally.

Measurement of total community density based on qPCR, however, do depend on a constant extraction efficiency and accurate measurement of sample quantity across samples. These assumptions may be systematically violated. In particular, methods dependent on the binding of DNA to a silica surface inevitably saturate [38, 39], where increasing input biomass does not result in proportional increases in extracted DNA [observed in e.g. 40]. The sensitivity of total yields to modifications of silica column-based steps suggests that this may affect standard extraction protocols [41]. In a study of five human fecal samples, approximately doubling extracted sample sizes from near 50mg to near 100mg resulted in less than proportional increases in estimated 16S rRNA gene abundance for all five samples (see Figure 4.2 panel A). If this reflects saturation of DNA extraction capacity, then the discrepancy is likely much larger at the kit recommended input mass of 250mg. While more comprehensive experimentation will be needed to understand these limitations, this observation suggests that qPCR-based community quantification does not necessarily reflect the true abundance of bacteria in the sample.

4.3 Spike-in quantification for studying microbial absolute abundance

Spike in quantification is carried out by adding known amounts of a recognizable DNA sequence to samples before extraction. The total size of the endogenous community is then estimated from the composition of sequences, leveraging the relationship between the fraction of the mixed community made up by the spike and the fraction of spike reads recovered after sequencing.

The formula to estimate, y_e, the absolute abundance of endogenous genes in the initial sample is

$$\widehat{y_e} = \lambda \frac{z_e y_s}{z_s} \tag{4.1}$$

where z_e and z_s are the read count for endogenous and spike genes, respectively,

and y_s is the known abundance of the spike. The term y_s can further broken down into the volume v_s and concentration c_s of marker genes in the spike. y_e is only rarely studied directly; instead, considering the density of bacteria in the original sample, $d_e = y_e/m_e$, enables comparison of samples with arbitrary differences in input sample weight.

The formula for estimating this density is therefore,

$$\hat{d}_e = \lambda \frac{z_e c_s v_s}{z_s m_e} \tag{4.2}$$

The calibration coefficient, λ, represents the proportional recovery of the spike genes relative to endogenous genes, accounting for biased extraction, amplification, and sequencing, which affect counts of each (see Figure 4.1). Since this coefficient is not known a priori—indeed, it may vary between experiments—it must be determined through calibration in order for values to be considered estimates of abundance. Alternatively, estimates of d_e/λ can be compared across samples and scale-invariant inferences about absolute abundance (e.g. "the absolute abundance is 2-fold higher") can still be made. Here, we refer to these uncalibrated measurements as "spike-adjusted density" since they are all scaled by the unknown λ^{-1}. Given a constant spike concentration, c_s can similarly be consolidated into the calibration coefficient, $\lambda' = \lambda c_s$, obviating the need for careful quantification of the spike.

This rearranged formula for spike-adjusted density is

$$\frac{\hat{d}_e}{\lambda'} = \frac{z_e v_s}{z_s m_e} \tag{4.3}$$

Formulas related to the above are used in Chapter 2, as well as by both Stämmler *et al.* [12] and Smets *et al.* [13].

Valid interpretation of $\hat{d}_e$ (as well as $d/\hat{\lambda}'$) depend on the calibration coefficient being constant across samples. To test this assumption for human fecal communities, spike-adjusted 16S gene abundance was compared to estimates obtained using qPCR. If λ differs by sample, little or no correlation between these two, independent measurements is expected. Instead, we find a tight relationship between log estimates (r = 0.96, see Figure 4.2 panel B) supporting the consistency of spike-in quantification.

These results reinforce similar, previously published findings in soil [12, 13]. While testing this under different experimental conditions is outside the scope of this chap-

ter, confirming the constancy of λ for new sample types, extraction kits, and choices of spike will strengthen future results using the approach. It is also unclear if comparisons of absolute abundance across multiple sample types are valid.

By adding the spike before extraction, spike-in quantification theoretically controls for variability in DNA extraction efficiency. However, the known saturation effects of silica column-based protocols raise the possibility that the recovery of spike and endogenous genes is differentially affected by changes in the amount of input sample used. This would result in a dependency of λ on m_e, violating the assumption that this coefficient is constant. We tested this possibility by comparing spike-adjusted density estimates across a wide range of input amounts (see Figure 4.2 panel C). While one sample did appear to have a monotonic relationship between the estimated density and input mass (Spearman's $\rho = 0.74$, $p = 0.001$), this relationship was driven by increased estimates in replicates with 25 mg input mass—much less than the protocol's specification—where physical heterogeneity of the sample might have impacted the accuracy of results. The four other samples showed no such relationship ($p > 0.05$).

4.3.1 Biological inference with absolute abundance data is non-trivial

While studies of community composition inherently limit the types of biological insights that can be obtained, the simplicity of generating relative abundance data has undoubtedly contributed to its widespread application. Several of the elegant features of compositional data are necessarily lost when scientific question require information about bacterial density.

One such feature: relative abundance is invariant to changes in sample mass. Whereas the quantity of sample being studied can be ignored in compositional studies, absolute abundance data is naturally interpreted in terms of population density, e.g. gene copies per gram of sample. This means that abundance estimates must be normalized to sample quantity, and measurement error in m_e directly contributes to error in the density estimate, $\widehat{d_e}$. What's more, the choice of units for sample quantity becomes an important decision. Studies of the gut microbiome, for instance, must now choose between per wet weight, per dry weight, or per volume bases for analyzing and reporting density. This is a non-trivial decision and has major implications for interpretation. Changes in material composition of the sample (e.g. hydration level, fiber content) will directly affect normalized measurements, and this may or may not

have biological importance. General recommendations for choice of basis are outside the scope of this chapter and depend heavily on the relevant biology.

To go from composition to the absolute abundance of individual taxa, relative abundance can be scaled by a measurement of the total community size to get an estimate for the actual abundance of individual taxa [11]. This has two important consequences that must be considered when analyzing these data. First, such estimates of absolute abundance necessarily lose precision since they compound noisy estimates of relative abundance with noisy estimates of absolute community density. Second, this latter source of noise is shared by all taxa in the sample. Consequently, correlations in the density of taxa will be artificially inflated, with noisier estimates resulting in stronger correlations. This confounder is circumvented by measuring taxa independently, most commonly with specifically designed qPCR primers [42]. Unfortunately, this means that untargeted, exploratory analyses are generally not possible, handicapping attempts to reconstruct complex ecological networks from correlation data.

Despite these limitations of all methods for the estimation of absolute abundance, spike-in quantification presents an attractive opportunity to introduce this valuable perspective to a broad range of studies in microbial ecology.

4.4 Optimizing spike-in quantification protocols

In this section factors affecting the accuracy and interpretability of spike-in quantification experiments are discussed. Guidelines and best-practices are proposed based on our experience working with human and mouse fecal samples. The major decisions in designing protocols revolve around the choice of what, when, and how much to spike.

4.4.1 What to spike

Previously published demonstrations of pre-extraction spike-in quantification have utilized intact bacterial cells as the spike. This means that the spike itself is exposed to the same extraction bias and variability as the endogenous community. Alternatively, a cell-free, DNA standard may be used, ensuring that spike recovery is independent of lysis efficiency. It is not yet clear which approach better maintains the constant proportionality between spike and endogenous sequence recovery—constant λ is the necessary feature for accurate quantification.

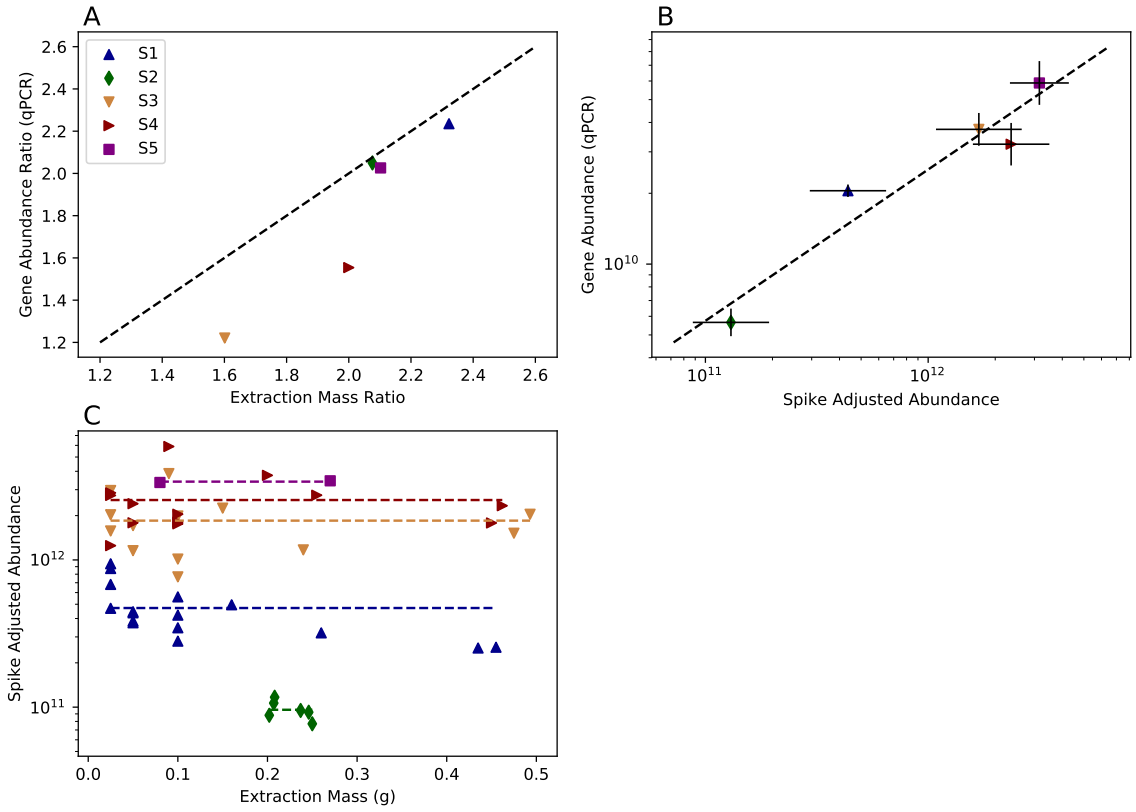

Figure 4.2: Accuracy and robustness of qPCR and spike-in quantification Accuracy and robustness of qPCR and spike-in quantification. (**A**) Comparison of sample mass ratios and 16S rRNA gene abundance estimates obtained using qPCR on DNA extractions from human feces. Points correspond with individual human fecal samples from five different subjects. Horizontal position of points indicates ratio of smaller and larger mass portions of same sample. Vertical position indicates the ratio in 16S rRNA gene abundance in final DNA extract, quantified using qPCR. The dotted line reflects the theoretical expectation of proportionality. Points under the line correspond to lower than expected estimates from larger fecal samples. (**B**) Comparison of estimated endogenous 16S rRNA gene density calculated using two independent methods. Points correspond with individual human fecal samples from five different subjects. Error bars represent standard deviations of log estimates from each method, and line of best fit (dashed) is shown. (**C**) Relationship between quantity of sample extracted and estimates of 16S rRNA gene density obtained from spike-in quantification. Points correspond with individual extractions, colored to indicate which fecal samples out of five, each from different human subjects. Dashed lines indicate the mean estimated gene density across all extractions for each sample.

In either case, the chosen spike sequence must be easily distinguished from all endogenous bacteria, while maintaining affinity for the PCR primers employed for amplification and sequencing. Cell-free DNA standards with appropriate primer binding sites can be constructed that differ from all known bacteria [43]. Alternatively, care must be taken with a cellular spike to ensure that sequences are not shared with endogenous members of the community. This means that no one cellular spike is appropriate for all possible experiments, although reasonably large subsets (e.g. all mammal associated microbiome studies) can share an appropriate standard.

For accurate comparison, the spike itself must also be kept identical for all samples. This standardization encompasses both the concentration of the spike reagent as well as the physical properties of the spike itself. This can be challenging for a cellular spike, as small variations in culture conditions can have a large impact on bacterial population density and physiology and therefore both gene density (due to genome copy number variation [44, 45]) and extraction efficiency (due to changes in cell wall physiology [46]). For a cellular spike, the use of a single batch of well-mixed bacterial culture is recommended to minimize variation between samples. Care must also be taken to ensure that cellular spikes are homogeneous over the volumetric scales being used, as large aggregates of cells can introduce substantial differences in the number of spike gene copies transferred. Standardizing the spike across studies has the potential to facilitate comparison, and a large culture volume can be aliquoted and stored for multiple experiments. Comparisons across studies using separately produced cellular spikes requires careful calibration for each. Standardization is easier when using a cell-free spike since DNA concentrations can be accurately measured [47].

Given the increasing application of metagenomic approaches, spiking in order to quantify total community size has implications for the simultaneous usability of DNA extractions for shotgun sequencing. While the addition of whole bacterial genomes means that these sequences also contaminate metagenomic libraries, if the spiked organism has been sequenced, reads can be removed by mapping to that reference. Alternatively, a cell-free, non-genomic spike greatly reduces the total number of spike reads in the metagenomic library. However, spike in approaches to quantification in metagenomics have also been explored [48], and the two can be integrated. In a dual analysis, we have observed accurate assembly of the genome of a cellular spike, lending confidence to other assemblies in the same study.

One option not covered in this work, but that has been previously explored [12], is the simultaneous use of multiple spike sequences. Studies could potentially benefit from features of both a cellular and cell-free spike. Dual spikes also provide an

internal check that recovery is proportional for each sequence across all samples. Analysis procedures for mixed spikes that take advantage of this internal control and improved precision have not yet been developed.

4.4.2 When to spike

Regardless of the choice of spike, choosing when to spike has important impacts on the accuracy and interpretation of quantification. While a cell-free spike can be added after DNA extraction, this limits the value of the approach in controlling for extraction variability. Indeed, spiking as early as possible in sample processing, even right at the time of sample collection, could present both logistical and experimental benefits. Spiking early improves quantification because both spike-volumes and sample quantities can be more accurately measured. Volumetric variation in aliquoting small quantities of fecal samples for extraction is a potentially large source of error, and individually weighing those aliquots can be logistically challenging.

Spiking early has two downsides: the contamination of potentially precious samples with foreign cells or DNA, and the need for larger quantities of spike reagent. We have not found these to be problematic in practice, but the trade-offs should be assessed on a per-study basis.

4.4.3 How much to spike

While the number of reads sequenced by high-throughput approaches has increased greatly in the past five years, this is still an important limitation on the number of samples included in community analysis studies, and also impacts the detection of rare taxa [49]. For this reason, the amount of spike added to each sample must be calibrated in order for an appropriate number of reads to be recovered. Adding too much spike can impact the recovery of rare taxa, or even swamp out the endogenous community entirely. On the other hand, recovering too little spike increases the impact of binomial sampling error on abundance estimates. At an extreme, when no reads of the spike are recovered for one or more samples it is impossible to use the Equation 4.2 to calculate $\hat{d}_e$.

As has been previously reported in molecular surveys [15], the number of spike reads recovered is overdispersed—more variable than would be predicted by an idealized binomial distribution. At higher counts, therefore, sampling error is dominated by overdispersion, and the optimal number of spike reads is not as dependent on the total number of reads sampled. Although not tested extensively, in our experience

recovering 100 spike reads has been sufficient, and increasing the spike beyond this point does not seem to improve the accuracy of estimation. In a library of 10,000 reads, this means that only 1% of the data is "lost" to spike reads. The challenge is ensuring approximately this number of reads when both library size and the density of endogenous bacteria vary. In our experience, uncertainty in library quantification prior to multiplex sequencing can result in approximately 2-fold differences in the number of reads per library. Aiming for additional spike reads can reduce the negative impacts on inference of this variation, while still wasting relatively little sequencing capacity. Much more challenging is the calibration of spike quantity in cases where endogenous bacterial densities may vary over multiple orders of magnitude, such as antibiotic treatment experiments [50–52]. When population sizes can be predicted before extraction, calibrating the amount of spike in order to recover approximately 100 reads is recommended. In other cases, first estimating densities in a subset of samples using qPCR or other methods will be valuable.

4.4.4 Analyzing sequence data from a spike-in quantification study

Once samples have been spiked, processing can proceed identically to standard community analysis up through digitization and demultiplexing of read libraries. Spike reads may then be counted and removed either before or after the standard bioinformatic pipeline. When spiked sequences are sufficiently similar to naturally occurring genes, counts can be partitioned directly from final, tabular outputs, since they are treated identically to endogenous taxa. This may not be the case, however, for synthetic sequences, which need only share primer binding sites with naturally occurring genes. Popular amplicon analysis pipelines, including MOTHUR [53] and QIIME [54], filter reads by length and alignment to reference databases. Therefore, synthetic sequences used as a spike may need to be counted and removed prior to other processing. Such spike reads will ideally be unambiguously identified by BLAST alignment to a reference sequence. Choosing alignment length and identity cutoffs to filter on with high sensitivity and specificity for the spike is an important consideration. Examining the impact of spike read processing on inferences is outside the scope of this manuscript.

4.5 Limitations in interpreting spike-in results

While the above guidelines for designing spike-in experiments enable fruitful application for quantification, in both our experience and in published studies [12, 13] measurement error with the approach can be large. However, even direct measurements may have similar levels of measurement error [11], and a variety of future improvements will undoubtedly increase the applicability and accuracy of the method. Measuring the calibration coefficient, λ, across sample types and protocol variations, as well as systematically optimizing the choice of spike and identifying one that can be obtained cheaply and reproducibly, will enable comparisons across studies.

Importantly, naïve analyses of spike-in quantification data are potential problematic. Given the ease with which point-estimates of absolute microbial density can be calculated from spike-in count data, traditional statistical tests operating directly on these point estimates are an attractive option for the analysis of such experiments. Point estimates are limited, however, by the discreteness and sparsity of count data. In particular, samples for which no reads of the spike are recovered are a critical challenge, since density estimates are therefore undefined. While such samples could be excluded from analysis, low spike recovery suggests high microbial abundance. Excluding such samples would therefore operate non-randomly, biasing statistical estimation.

One option is to use pseudocounts, adding 1 to every cell in the count matrix, prior to calculating point estimates. This has two effects. First, absolute abundance estimates based on pseudocounts are defined and non-zero for all taxa in all samples. Second, pseudocounts implicitly bias abundance estimates. Fortunately, samples with high spike recovery are only minimally affected, reflecting their greater theoretical precision.

A statistical approach designed to natively handle the challenging features of this data type is described in Chapter 5.

4.6 Conclusions

In this chapter, the utility of spike-in quantification in microbial community analysis has been described. We believe that the approach should be considered in a substantial fraction of future studies. While qPCR and other direct quantification approaches are also valuable, the simplicity and low cost of spike-in quantification enables examination of absolute abundance in studies where it has previously been

ignored.

A breadth of experimental considerations have been discussed that impact accurate and reproducible use of spike-in quantification, and a number of best practices for choosing a spike, designing protocols, and analyzing the resulting sequence data have been recommended.

4.7 Methods

4.7.1 Sample collection

Two fecal samples from each of five separate individuals were collected using the OMNIgene-GUT microbiome sampling system (DNAGenotek). Tubes containing collection buffer were weighed before and after sampling. Paired samples were mixed well and combined into a single slurry per-individual. Aliquots of these slurries reserved for qPCR-based quantification were frozen at -20 before further processing.

4.7.2 Extraction and qPCR for direct quantification

For each slurry, DNA from weighed subsamples of approximately 50 and 100 mg was extracted using the DNeasy PowerSoil Pro (Qiagen Ref 47104) DNA extraction kit, with final elution into 50 μLof water. The abundance of 16S rRNA genes in these DNA samples was quantified by qPCR with the Qiagen QuantiTect SYBR Green MasterMix on the Roche LightCycler 96 platform. Terminal melt curves between 65 and 97 were used for validation. *Escherichia coli* genomic DNA (Sigma D4889) was used as a quantification standard and positive control.

4.7.3 Spike reagent

All spikes were done with a raw DNA spike prepared by the Institute for Life Science Entrepreneurs. This DNA construct is 372 bp long with concatenated standard 16S rRNA gene primer binding sites on either side, of a 48 bp synthetic "ID Tag" sequence. Two distinct constructs were obtained from ILSE at 1×10^8 copies / μLand combined in equal proportions to a final concentration of 5×10^7 copies / μLof each.

4.7.4 Demo experiments

Demonstration spike-in quantification experiments were performed on all five human fecal samples, using one or more of three slightly different procedures. For "Series

A" approximately 2.6 mL of samples S1, S3, and S4 were transferred to fresh tubes, weighed, and 26 μLof the mixed spike reagent was added to each. These slurries were then aliquoted into the extraction plate in one or more replicates at approximately 400 μL, 200 μL, 100 μL, 50 μL, and 25 μL. Transfers greater than 100 μLwere directly weighed to account for pipetting error. For smaller volumes, weights were estimated from an estimate of slurry density. For "Series B", approximately 1 mL of S2 and S5 were transferred to fresh tubes, weighed, and 10 μLof mixed spike reagent was added to each. Approximately 100 μLof this mixture was then aliquoted into the extraction plate with at least 4 replicates of each. For "Series C", remaining unspiked slurry from samples S1 through S5 were aliquoted in individually weighed approximately 200 μLincrements into the extraction plate.

Where appropriate, results from all three procedures were pooled together for analysis.

4.7.5 Extraction, sequencing, and bioinformatic processing

Extraction was performed with the Qiagen PowerMag Microbiome kit. The V4 hypervariable region of the 16S rRNA gene was amplified as described in Chapter 2. Amplicons were then sequenced on an Illumina MiSeq using MiSeq Reagent Kit V2 500 cycles.

Paired-end sequences were fused using MOTHUR [53]. Spike reads were identified with BLAST [55] against their reference sequences using a 95% identity threshold, minimum alignment length of 150 positions, and a maximum alignment length of 300. Reads from both spike sequences were combined into a single spike tally. Read libraries were processed using the MOTHUR pipeline based on the 16S standard operating procedures [56].

BIBLIOGRAPHY

[1] Schmidt, T.M., DeLong, E.F., Pace, N.R.N., Biology, C.: Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. Journal of bacteriology **173**(14), 4371–4378 (1991). doi:10.1128/jb.173.14.4371-4378.1991

[2] Weir, T.L., Manter, D.K., Sheflin, A.M., Barnett, B.A., Heuberger, A.L.: Stool Microbiome and Metabolome Differences between Colorectal Cancer Patients and Healthy Adults. PLoS ONE **8**(8), 70803 (2013). doi:10.1371/journal.pone.0070803

[3] Ahn, J., Sinha, R., Pei, Z., Dominianni, C., Wu, J., Shi, J., Goedert, J.J., Hayes, R.B., Yang, L.: Human Gut Microbiome and Risk of Colorectal Cancer. Journal of the National Cancer Institute **105**(24), 1907–1911 (2013). doi:10.1093/jnci/djt300

[4] Turnbaugh, P.J., Ley, R.E., Mahowald, M.A.: An obesity-associated gut microbiome with increased capacity for energy harvest. Nature **444** (2006)

[5] Fahlén, A., Engstrand, L., Baker, B.S., Powles, A., Fry, L.: Comparison of bacterial microbiota in skin biopsies from normal and psoriatic skin. Archives of Dermatological Research **304**(1), 15–22 (2012). doi:10.1007/s00403-011-1189-x

[6] Sampson, T.R., Mazmanian, S.K.: Control of brain development, function, and behavior by the microbiome. Cell Host and Microbe **17**(5), 565–576 (2015). doi:10.1016/j.chom.2015.04.011. 15334406

[7] Tsilimigras, M.C.B., Fodor, A.A.: Compositional data analysis of the microbiome: fundamentals, tools, and challenges. Elsevier Inc (2016). doi:10.1016/j.annepidem.2016.03.002. http://dx.doi.org/10.1016/j.annepidem.2016.03.002

[8] Aitchison, J., J. Egozcue, J.: Compositional Data Analysis: Where Are We and Where Should We Be Heading? Mathematical Geology **37**(7), 829–850 (2005). doi:10.1007/s11004-005-7383-7

[9] Faith, J.J., Faith, J.J., Mcnulty, N.P., Rey, F.E., Gordon, J.I.: Response to Diet in Gnotobiotic Mice **101**(2011), 101–105 (2013). doi:10.1126/science.1206025

[10] Blagodatsky, S., Blagodatskaya, E., Yuyukina, T., Kuzyakov, Y.: Model of apparent and real priming effects: Linking microbial activity with soil organic matter decomposition. Soil Biology and Biochemistry **42**(8), 1275–1283 (2010). doi:10.1016/j.soilbio.2010.04.005

[11] Zhang, Z., Qu, Y., Li, S., Feng, K., Wang, S., Cai, W., Liang, Y., Li, H., Xu, M., Yin, H., Deng, Y.: Soil bacterial quantification approaches coupling with relative abundances reflecting the changes of taxa. Scientific Reports **7**(1), 4837 (2017). doi:10.1038/s41598-017-05260-w

[12] Stämmler, F., Gläsner, J., Hiergeist, A., Holler, E., Weber, D., Oefner, P.J., Gessner, A., Spang, R.: Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. Microbiome, 1–13 (2016). doi:10.1186/s40168-016-0175-0

[13] Smets, W., Leff, J.W., Bradford, M.A., McCulley, R.L., Lebeer, S., Fierer, N.: A method for simultaneous measurement of soil bacterial abundances and community composition via 16S rRNA gene sequencing. Soil Biology and Biochemistry **96**, 145–151 (2016). doi:10.1016/j.soilbio.2016.02.003

[14] Morton, J.T., Sanders, J., Quinn, R.A., Mcdonald, D., Gonzalez, A., Vázquez-baeza, Y., Navas-molina, J.A.: Balance trees reveal microbial niche differentiation **2**(1), 1–11 (2017). doi:10.1128/mSystems.00162-16

[15] McMurdie, P.J., Holmes, S.: Waste not, want not: why rarefying microbiome data is inadmissible. PLoS computational biology **10**(4), 1003531 (2014). doi:10.1371/journal.pcbi.1003531

[16] Ubeda, C., Taur, Y., Jenq, R.R., Equinda, M.J., Son, T., Samstein, M., Viale, A., Socci, N.D., Van Den Brink, M.R.M., Kamboj, M., Pamer, E.G.: Vancomycin-resistant Enterococcus domination of intestinal microbiota is enabled by antibiotic treatment in mice and precedes bloodstream invasion in humans. Journal of Clinical Investigation **120**(12), 4332–4341 (2010). doi:10.1172/JCI43918

[17] Everard, A., Lazarevic, V., Derrien, M., Girard, M., Muccioli, G.M., Neyrinck, A.M., Possemiers, S., Van Holle, A., François, P., De Vos, W.M., Delzenne, N.M., Schrenzel, J., Cani, P.D.: Responses of gut microbiota and glucose and lipid metabolism to prebiotics in genetic obese and diet-induced leptin-resistant mice. Diabetes **60**(11), 2775–2786 (2011). doi:10.2337/db11-0227

[18] Fisher, C.K., Mehta, P.: Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. PLoS ONE **9**(7), 1–10 (2014). doi:10.1371/journal.pone.0102451. 1402.0511

[19] Deng, Y., Jiang, Y., Yang, Y., He, Z., Luo, F., Zhou, J.: Molecular ecological network analyses. BMC bioinformatics (2012)

[20] Faust, K., Sathirapongsasuti, J.F., Izard, J., Segata, N., Gevers, D., Raes, J., Huttenhower, C.: Microbial co-occurrence relationships in the human microbiome. PLoS Computational Biology **8**(7), 1002606 (2012). doi:10.1371/journal.pcbi.1002606

[21] Kurtz, Z.D., Müller, C.L., Miraldi, E.R., Littman, D.R., Blaser, M.J., Bonneau, R.A.: Sparse and Compositionally Robust Inference of Microbial Ecological Networks. PLOS Computational Biology **11**(5), 1004226 (2015). doi:10.1371/journal.pcbi.1004226

[22] Biswas, S., Mcdonald, M., Lundberg, D.S., Dangl, J.L., Jojic, V.: Learning Microbial Interaction Networks from Metagenomic Count Data. Journal of Computational Biology **23**(6), 526–535 (2016). doi:10.1089/cmb.2016.0061

[23] Props, R., Kerckhof, F.M., Rubbens, P., Vrieze, J.D., Sanabria, E.H., Waegeman, W., Monsieurs, P., Hammes, F., Boon, N.: Absolute quantification of microbial taxon abundances. ISME Journal **11**(2), 584–587 (2017). doi:10.1038/ismej.2016.117

[24] Vandeputte, D., Kathagen, G., D'Hoe, K., Vieira-Silva, S., Valles-Colomer, M., Sabino, J., Wang, J., Tito, R.Y., De Commer, L., Darzi, Y., Vermeire, S., Falony,

G., Raes, J.: Quantitative microbiome profiling links gut community variation to microbial load. Nature **551**(7681), 507–511 (2017). doi:10.1038/nature24460

[25] Abzazou, T., Salvadó, H., Bruguera-Casamada, C., Simón, P., Lardín, C., Araujo, R.M.: Assessment of total bacterial cells in extended aeration activated sludge plants using flow cytometry as a microbial monitoring tool. Environmental Science and Pollution Research **22**(15), 11446–11455 (2015). doi:10.1007/s11356-015-4372-3

[26] Frossard, A., Hammes, F., Gessner, M.O.: Flow cytometric assessment of bacterial abundance in soils, sediments and sludge. Frontiers in Microbiology **7**(JUN), 1–8 (2016). doi:10.3389/fmicb.2016.00903

[27] Smith, C.J., Osborn, A.M.: Advantages and limitations of quantitative PCR (Q-PCR)-based approaches in microbial ecology (2009). doi:10.1111/j.1574-6941.2008.00629.x

[28] Davis, C.: Enumeration of probiotic strains: Review of culture-dependent and alternative techniques to quantify viable bacteria. Journal of Microbiological Methods **103**, 9–17 (2014). doi:10.1016/j.mimet.2014.04.012

[29] Frostegård, Å., Bååth, E., Tunlio, A.: Shifts in the structure of soil microbial communities in limed forests as revealed by phospholipid fatty acid analysis. Soil Biology and Biochemistry **25**(6), 723–730 (1993). doi:10.1016/0038-0717(93)90113-P

[30] Hammes, F., Goldschmidt, F., Vital, M., Wang, Y., Egli, T.: Measurement and interpretation of microbial adenosine tri-phosphate (ATP) in aquatic environments. Water Research **44**(13), 3915–3923 (2010). doi:10.1016/j.watres.2010.04.015

[31] Frostegård, A., Bååth, E.: The use of phospholipid fatty acid analysis to estimate bacterial and fungal biomass in soil. Biology and Fertility of Soils **22**(1-2), 59–65 (1996). doi:10.1007/BF00384433

[32] Holm-Hansen, O.: Determination of Microbial Biomass in Ocean Profiles. Limnology and Oceanography **14**(4), 740–747 (1969). doi:10.4319/lo.1969.14.5.0740

[33] Blagodatskaya, E.V., Blagodatsky, S.A., Anderson, T.H., Kuzyakov, Y.: Priming effects in Chernozem induced by glucose and N in relation to microbial growth strategies. Applied Soil Ecology **37**(1-2), 95–105 (2007). doi:10.1016/j.apsoil.2007.05.002

[34] Buchkowski, R.W., Schmitz, O.J., Bradford, M.A.: Microbial stoichiometry overrides biomass as a regulator of soil carbon and nitrogen cycling. Ecology **96**(4), 1139–1149 (2015). doi:10.1890/14-1327.1

[35] Henderson, G., Cox, F., Kittelmann, S., Miri, V.H., Zethof, M., Noel, S.J., Waghorn, G.C., Janssen, P.H.: Effect of DNA extraction methods and sampling techniques on the apparent structure of cow and sheep rumen microbial communities. PloS one **8**(9), 1–14 (2013). doi:10.1371/journal.pone.0074787

[36] Feinstein, L.M., Sul, W.J., Blackwood, C.B.: Assessment of bias associated with incomplete extraction of microbial DNA from soil. Applied and environmental microbiology **75**(16), 5428–33 (2009). doi:10.1128/AEM.00120-09

[37] Stults, J.R., Snoeyenbos-West, O., Methe, B., Lovley, D.R., Chandler, D.P.: Application of the 5' Fluorogenic Exonuclease Assay (TaqMan) for Quantitative Ribosomal DNA and rRNA Analysis in Sediments. Applied and Environmental Microbiology **67**(6), 2781–2789 (2001). doi:10.1128/AEM.67.6.2781-2789.2001

[38] Poeckh, T., Lopez, S., Fuller, A.O., Solomon, M.J., Larson, R.G.: Adsorption and elution characteristics of nucleic acids on silica surfaces and their use in designing a miniaturized purification unit. Analytical Biochemistry **373**(2), 253–262 (2008). doi:10.1016/j.ab.2007.10.026. NIHMS150003

[39] Melzak, K.A., Sherwood, C.S., Turner, R.F.B., Haynes, C.A.: Driving forces for DNA adsorption to silica in perchlorate solutions. Journal of Colloid and Interface Science **181**(2), 635–644 (1996). doi:10.1006/jcis.1996.0421

[40] Katevatis, C., Fan, A., Klapperich, C.M.: Low concentration DNA extraction and recovery using a silica solid phase. PLoS ONE **12**(5), 1–14 (2017). doi:10.1371/journal.pone.0176848

[41] Miller, D.N., Bryant, J.E., Madsen, E.L., Ghiorse, W.C., Al, M.E.T.: Evaluation and optimization of DNA extraction and purification procedures for soil and sediment samples. Applied and Environmental Microbiology **65**(11), 4715–4724 (1999). doi:10.1016/0003-2697(80)90589-8

[42] Rinttilä, T., Kassinen, A., Malinen, E., Krogius, L., Palva, A.: Development of an extensive set of 16S rDNA-targeted primers for quantification of pathogenic and indigenous bacteria in faecal samples by real-time PCR. Journal of Applied Microbiology **97**(6), 1166–1177 (2004). doi:10.1111/j.1365-2672.2004.02409.x

[43] Roy, S., Caruthers, M.: Synthesis of DNA/RNA and their analogs via phosphoramidite and H-phosphonate chemistries. Molecules **18**(11), 14268–14284 (2013). doi:10.3390/molecules181114268

[44] Schaechter, M., MaalOe, O., Kjeldgaard, N.O.: Dependency on Medium and Temperature of Cell Size and Chemical Composition during Balanced Growth of Salmonella typhimurium. Journal of General Microbiology **19**(3), 592–606 (1958). doi:10.1099/00221287-19-3-592

[45] Dennis, P.P., Bremer, H.: Modulation of Chemical Composition and Other Parameters of the Cell at Different Exponential Growth Rates. EcoSal Plus **3**(1) (2008). doi:10.1128/ecosal.5.2.3

[46] Makinoshima, H., Aizawa, S.I., Hayashi, H., Miki, T., Nishimura, A., Ishihama, A.: Growth phase-coupled alterations in cell structure and function of Escherichia coli. Journal of Bacteriology **185**(4), 1338–1345 (2003). doi:10.1128/JB.185.4.1338-1345.2003

[47] Rengarajan, K., Cristol, S.M., Mehta, M., Nickerson, J.M.: Quantifying DNA concentrations using fluorometry: A comparison of fluorophores. Molecular Vision **8**(November), 416–421 (2002)

[48] Satinsky, B.M., Gifford, S.M., Crump, B.C., Moran, M.A.: Use of internal standards for quantitative metatranscriptome and metagenome analysis. Methods in Enzymology **531**, 237–250 (2013). doi:10.1016/B978-0-12-407863-5.00012-5

[49] Hamp, T.J., Jones, W.J., Fodor, A.A.: Effects of experimental choices and analysis noise on surveys of the "Rare Biosphere". Applied and Environmental Microbiology **75**(10), 3263–3270 (2009). doi:10.1128/AEM.01931-08

[50] Jernberg, C., Löfmark, S., Edlund, C., Jansson, J.K.: Long-term impacts of antibiotic exposure on the human intestinal microbiota. Microbiology **156**(11), 3216–3223 (2010). doi:10.1099/mic.0.040618-0

[51] Rashid, M.U., Weintraub, A., Nord, C.E.: Effect of new antimicrobial agents on the ecological balance of human microflora. Anaerobe **18**(2), 249–253 (2012). doi:10.1016/j.anaerobe.2011.11.005

[52] Sullivan, Å., Edlund, C., Nord, C.E.: Effect of antimicrobial agents on the ecological balance of human microflora. The Lancet Infectious Diseases **1**(2), 101–114 (2001). doi:10.1016/S1473-3099(01)00066-4

[53] Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., Sahl, J.W., Stres, B., Thallinger, G.G., Van Horn, D.J., Weber, C.F.: Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology **75**(23), 7537–7541 (2009). doi:10.1128/AEM.01541-09

[54] Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.a., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.a., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.a., Widmann, J., Yatsunenko, T., Zaneveld, J., Knight, R.: QIIME allows analysis of high-throughput community sequencing data. Nature Methods **7**(5), 335–336 (2010). doi:10.1038/nmeth.f.303. NIHMS150003

[55] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. Journal of Molecular Biology **215**(3), 403–410 (1990). doi:10.1016/S0022-2836(05)80360-2. arXiv:1611.08307v1

[56] Kozich, J.J., Westcott, S.L., Baxter, N.T., Highlander, S.K., Schloss, P.D.: Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the miseq illumina sequencing platform. Applied and Environmental Microbiology **79**(17), 5112–5120 (2013). doi:10.1128/AEM.01043-13

CHAPTER 5

A novel, model-based approach for inference on microbial absolute abundance leveraging spike-in quantification data

5.1 Background

Community surveys harnessing the 16S rRNA gene have been the central tool in microbial ecology for more than a decade [1], enabling the discovery of associations between bacterial composition and experimental or observed covariates, and has driven exciting discoveries related to the human microbiome [2–4] and other fields. However, traditional community surveys are unable to conclusively demonstrate changes in absolute abundance of individual bacterial taxa, and analyses must instead depend on compositional, relative abundance data. Microbial ecology has adapted to this shortcoming by adopting appropriate statistical tools [5–12]. Unfortunately, some of the key inferential goals of community ecology are inaccessible without absolute abundance information, in particular predicting metabolic impact and characterizing interspecific interactions [13].

Despite the inherent limitations of relative abundance data in the analysis of microbial communities, methods intended to directly measure absolute abundance, such as qPCR, have found only limited use. This may be because of technical challenges in accurately and efficiently scaling these protocols to the potentially hundreds of samples considered in sufficiently powered studies.

Recently, spike-in quantification has been proposed as an alternative approach [14, 15], enabling simple estimation of the abundance of taxonomic marker genes serving as a proxy for the absolute abundance of microbes. In this protocol, a known amount of an identifiable marker gene sequence is added to samples before processing. After collecting sequence data from the mixed sample, the gene density of the endogenous

microbial community, d, may then be estimated with the formula

$$\hat{d} = \lambda \frac{z_e v_s c_s}{z_s m_e} \tag{5.1}$$

where z_s and z_e is the read count for the spike sequence and endogenous sequences, respectively, v_s, and c_s are the known volume and concentration of spike sequence added, and m_e is the known mass of the sample being analyzed. The density of individual members of that community can then be estimated by scaling relative abundance by $\hat{d}$.

While under an idealized scenario the estimate of d becomes precise as total read counts increase, for finite library sizes the discreteness of counts and binomial error of sampling z_e and z_s introduce noise. In addition, excess error introduced by sample heterogeneity, and the variability of extraction, PCR, and sequencing, reduces the accuracy of these point estimates. Statistical analyses that operate directly on $\hat{d}$, therefore, fail to appropriately propagate this uncertainty. In addition, measurement error likely deviates from the distributional assumptions implicit in common statistical procedures such as ANOVA and linear regression.

In this paper an integrated model for microbial abundance and spike-in quantification is developed and applied to the analysis of experimental data, leveraging this recently demonstrated protocol for improved inference of microbial community dynamics. In Section 5.2 a probabilistic model is described for the abundance of microbes as well as count data resulting from spike-in quantification of multi-species samples. In Section 5.3 a software tool, SpikeAbund, is introduced for fitting this model to real data and to enable inference on model parameters under a Bayesian framework. In Section 5.4 simulated data is used to compare the performance of SpikeAbund to naïve analysis methods. In Section 5.5, SpikeAbund is applied to the analysis of real data in order to identify bacteria residing in the guts of mice that are affected by treatment with the α-glucosidase inhibitor acarbose. Finally, in Section 5.6, we criticize the model using these data, and consider an extension that may better reflect reality.

Spike-in quantification has the potential to better align marker gene surveys with biologically relevant features of microbial communities. Given the complexity of the resulting data and the potential for deviations from distributional assumptions, tailored statistical procedures are needed to enable correct and efficient interpretation. This model-based approach to inference leverages spike-in quantification for novel

insights, and presents a platform for further exploration of microbial community dynamics.

5.2 A statistical model of community abundance and spike-in quantification

In this section a plausible data generating process is introduced for count data from a spike-in quantification procedure on samples from experimental or observational studies in microbial systems. Potential modifications of the model which may better describe reality are also discussed.

As with other sequence-based surveys of microbial communities, the raw output from a spike-in quantification experiment is multivariate count data, which may contain zeros, and is implicitly correlated due to compositionality. The Dirichlet-multinomial (DM) distribution [16] has been successfully used to model such data both in microbial [17–19] and other ecology [6] and in numerous other fields [20]. Due to its relationship with the multinomial distribution, the DM natively accommodates properties of this data type, while accounting for overdispersion as a Dirichlet mixture [21].

The DM is used to describe the distribution of counts for k sequences and the spike. For ease of interpretation, here we parameterize the underlying Dirichlet with a vector Π on the k-simplex, which defines the expectation for each element, and a scalar α determining the concentration. The more commonly used parameterization is obtained from the product of these two. As α increases, the DM approaches a multinomial distribution; finite values of α therefore describe the overdispersion (or "clumpiness") of real, biological data, which in amplicon libraries may be due to sample heterogeneity, extraction variability, and PCR dynamics.

The fraction of each read is then specified by Π a deterministic transformation of the latent abundance for each sequence, y_j, accounting for the addition of a known quantity of spike sequence, s. Here, an independently distributed, normal, linear model for the log-abundance of each taxon is used. The effects of covariates x_1 through x_K on taxon log-abundance are described by the parameters β_1 through β_K with baseline log abundance β_0.

The full likelihood model for sample i:

$$
\begin{aligned}
Z_i &\sim \mathrm{DM}(\alpha\Pi_i, m) \\
\Pi_i &= \left[\frac{y_{i1}}{s_i + d_i}, \ldots, \frac{y_{iJ}}{s_i + d_i}, s_i\right] \\
d_i &= \sum_{j=1}^{J} y_{ij} \\
\log(y_{ij}) &\sim \mathrm{Norm}(\mu_{ij}, \sigma_j) \\
\mu_{ij} &= \beta_{0j} + \sum_{k=1}^{K} x_{ik}\beta_{kj}
\end{aligned}
$$

With read counts and a known abundance of spike in each sample, latent taxon abundances, y_{ij}, are constrained. This enables estimation of model parameters. Of particular interest are the values of β_{kj} which represent the effect of covariate k on the abundance of taxon j.

Several elements of this model are obvious candidates for modification when biological reality deviates. In particular, the linear model for mean log-abundance allows for latent abundances with no upper limit or saturation. Likewise, the multiplicative noise means that highly abundant organisms may fluctuate to still higher abundances. These features are implausible in microbial communities that are, at a minimum, subjects to spatial constraints on maximum population size. Functional relationships between covariates and abundance that better describe reality, in particular those that account for this necessary saturation, may be provide an improved description of reality.

The model proposed here also ignores interactions between taxa by drawing abundances from independent distributions. At the simplest extreme, introducing covariance between taxa by modeling latent abundances as draws from a multivariate normal distribution may sufficiently approximate this biology. In addition, the DM distribution constrains the stochastic processes generating count data from underlying abundances; for example, elements with the same mean always have the same variance. Replacement of the sampling model with more general covariance structures, for example the generalized Dirichlet multinomial [21], could better fit observed counts.

In the following sections we describe the use of this model for the analysis of both simulated and real data. An iterative process of building and criticizing models with available data has the potential to improve our understanding of microbial communities. By directly modeling spike-in quantification results, new insights can be

obtained into microbial community dynamics that are not available in compositional datasets.

5.3 SpikeAbund is a command-line tool for the analysis of spike-in quantification data

By leveraging this integrated model of spike-in quantification for the analysis of data, the values of various parameters may be interrogated. A Bayesian framework enables intuitive interpretation, while natively handling the large number of latent parameters in the model. This approach to spike-in data analysis is packaged as the software tool SpikeAbund that can be run on Linux, Windows, or macOS. Code and documentation are made freely available.

5.3.1 Inputs

SpikeAbund takes a matrix of taxonomic marker gene counts that can be produced by a variety of widely used amplicon library analysis software, including MOTHUR [22] and QIIME [23]. A second, metadata file is also required, with the values of observed and experimentally controlled covariates. This file should also include the normalized spike-in amount for each sample, representing $v_s m_e^{-1}$ (see Formula 5.1). Normalization of the spike input to sample mass or volume is crucial for interpretation the model in terms of microbial population density, and is appropriate for most uses. If users have calibrated their spike protocol, $\lambda v_s c_s m_e^{-1}$ may be used instead, and the values of β_0 can be interpreted as the true gene density, rather than spike-adjusted densities. This will not affect other β terms.

Users may also specify the linear model relating covariates to abundance, and identify which column of the counts table represents the spike sequence. The primary output of the program is a summary of the marginal posterior distribution for each model parameter. A full MCMC chain can also be saved for downstream analysis. Advanced visualization and inference may be carried out in the fully interactive IPython environment by making use of `%run` magic.

5.3.2 Priors

Priors were selected to be weakly informative. A normally distributed prior was chosen for the log of the concentration parameter, α, with a standard deviation of 10. A

normal distribution was also selected, with mean 0 and (by default) a standard deviation of 10, for β. For binary covariates, this prior corresponds with approximately 68% of abundance effects being within an approximately 22000-fold change, keeping parameters within a manageable range even in cases where data does not constrain belief, but having a negligible effect when sufficient data is available. This prior becomes more influential as the magnitude of the covariate is increased. Standardization of covariates is therefore recommended. Alternatively, the standard deviation of this normal prior may be increased to lessen the potential for biased inference, or may be decreased, implying skepticism of large effect sizes and regularizing parameter estimates. Checking the sensitivity of inference to the value of this user-defined parameter is recommended. Future versions of SpikeAbund will explore alternative prior distributions.

5.3.3 Implementation

Bayesian inference of parameters based on this approximate model is made available as a software tool written in python and heavily utilizing the PyMC3 package for probabilistic programming [24]. The linear model relating covariates to the expected log-abundance of each organism may be specified using R-like formula notation as implemented by the python package Patsy [25]. The No-U-Turn Sampler [26] is used for efficient Markov chain Monte Carlo (MCMC) sampling from the posterior distribution, and inference is performed directly on this posterior samples.

5.3.4 MCMC convergence and quality diagnostics

Before trusting a simulated posterior, It is important to confirm that MCMC sampling converges to the target distribution, and that the Markov chain is well-behaved. The PyMC3 packages implements and automatically performs several standard checks: flagging potential divergences in the Hamiltonian Monte Carlo error during sampling [27], reporting an estimate of the effective number of samples from the posterior based on the autocorrelation of the chain, and also reporting the Gelman-Rubin diagnostic [28], an indicator of failed convergence. The user is warned when normal thresholds are exceeded. While not exhaustive, we have not observed cases of failed sampling where these diagnostics were unable to detect the problem. When these quality checks are failed, the user has the option to adjust parameters in order to improve sampling. Divergences during sampling can sometimes be resolved by increase the target accept probability (0.8 by default). Failure to converge can be fixed with

additional tuning steps (500 by default). And poor mixing can be overcome with increased chain lengths (1000 by default) or, alternatively, additional chains may be sampled in parallel (2 by default).

5.3.5 Interpreting posterior distributions

As with other Bayesian approaches to data analysis, rich interpretations may be obtained from the posterior distribution. In particular, this software reports mean parameter values as well as 50% and 95% highest-probability density credible intervals, by default.

In order to provide a widely accessible criterion for reporting the effects of individual covariates on individual taxa—a role often played by *P*-values in frequentist approaches—the posterior probability that the true parameter value has the opposite sign from the reported mean, known as a type-S error [29, 30], is reported. Users may instead wish to quantify the posterior probability that a covariate has biological relevance, defined as some threshold effect size. Posterior distributions may be applied in this way, representing a key theoretic advantage over frequentist approaches.

5.4 Comparison of analysis methods on simulated data

A standard protocol for the analysis of spike-in quantification data for microbial communities has not yet been established. Here SpikeAbund is compared to two naïve procedures: non-parametric comparison of rank abundance using a Mann-Whitney U test (here abbreviated as sMWU), and comparison of mean log-abundance using a t-test (sTT). Furthermore, in order to demonstrate the value of spike-in quantification, two parallel procedures are included on compositional data comparing rank relative abundance (cMWU) and mean log relative abundance (cTT). The performance of these procedures is compared on simulations of a relevant experimental scenario. Unsurprisingly, spike-adjusted abundance outperforms relative abundance for detecting changes in population size. Evidence is also provided that a model-based approach to analysis of spike-in quantification data supersedes the other methods both conceptually and in practice.

In most sequence-based studies of microbial communities, two goals of statistical analysis are to identify taxa that are differentially abundant among treatments and to estimate the direction and magnitude of that effect. The first goal is more

Table 5.1: Summary of methods for the analysis of spike-in quantification data

Abbreviation	Description	Binary inference	Parameter estimation
cMWU[a]	MWU test on observed composition	*P*-value	n/a
cTT[a]	t-test on log observed composition	*P*-value	Difference in mean log relative abundance
sMWU[b]	MWU test on spike-adjusted abundance	*P*-value	n/a
sTT[b]	t-test on log spike-adjusted abundance	*P*-value	Difference in mean log spike-adjusted abundance
SpikeAbund[b]	Dirichlet multinomial spike-in counts model	Posterior probability of type-S error	Mean of the posterior distribution of β

[a] Interpretation based on relative abundance
[b] Interpretation based on absolute abundance

commonly achieved using statistical tests and a *P*-value cutoff. Effects can then be calculated from estimates of central tendency, such as differences in mean or median. Alternatively, for more complicated models, regression-based approaches are often used.

In order to compare methods for analyzing microbial communities, count data were simulated from a hypothetical experiment in which the effect of a polysaccharide prebiotic on gut bacteria is tested. Scenarios like this one may be a particularly misleading when analyzing relative abundance data. Given the taxon specific, but generally positive effect of such a treatment on bacterial populations, increases in the absolute abundance of some taxa can obfuscate changes in others. Treatments such as antibiotics that decrease total community size should have similar impacts on analysis.

Performance is compared across the simulated taxa. For clarity, we characterize these based on the expected number of reads in control samples out of 5,000 simulated reads: high abundance taxa (Ranks 1 through 14) with greater than 100, moderately abundant taxa (Rank 15 through 30) with more than 10, low abundance taxa (Ranks 31 through 47) with more than 1, and very low abundance taxa (Ranks 48 and 49) with expected read counts of less than 1.

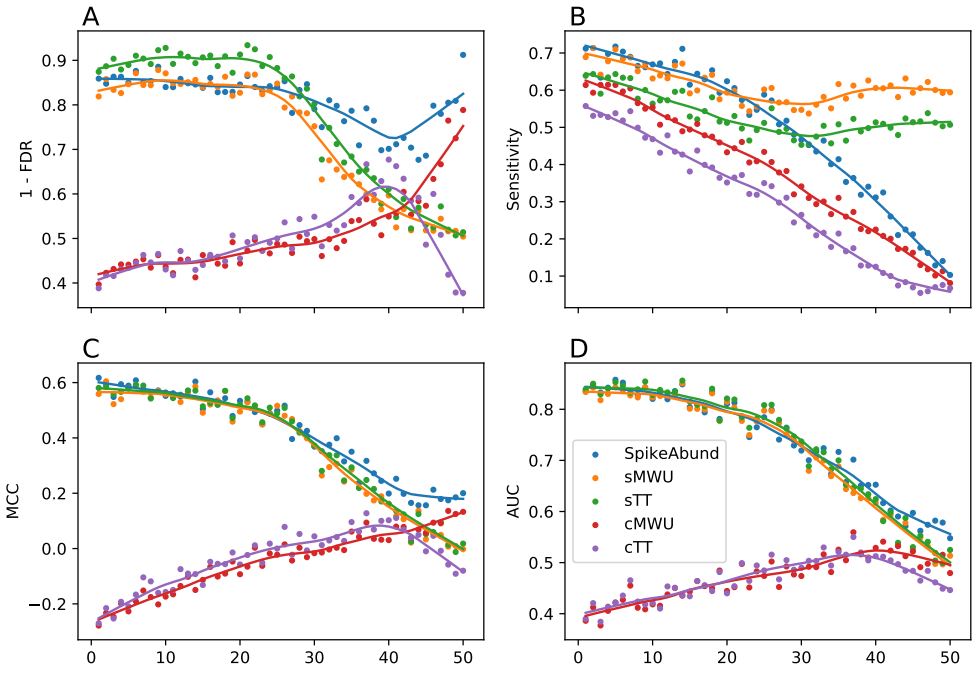

Figure 5.1: Comparison of classification skill across five procedures for the analysis of changes in taxon abundance. Skill is reported as 1 - false discovery rate (**A**), sensitivity (**B**), Matthew's correlation coefficient (**C**) and area under the receiver operating characteristic curve (**D**). Points are the skill of the procedure for each taxon rank (decreasing abundance) across 1,000 simulations. Higher values are better. Colors correspond to analysis procedure, with three utilizing spike-in quantification, and two operating directly on relative abundance. Trends are visually summarized with lowess regression lines.

Unsurprisingly, procedures designed to find differences in relative abundance failed to accurately identify real changes in the underlying absolute abundance. For abundant taxa, false discovery rates were greater than 50%. Given the simulated distribution of true effects, this represents worse performance than the expectation for random guessing; Matthew's correlation coefficients (MCCs) are consequently less than zero for these taxa. This poor performance reflects the mismatch between relative and absolute abundance. In the very low abundance taxa, cMWU has improved performance: slightly better than random guessing. To confirm that the performance of these tests could not be improved by choosing a different *P*-value cutoff, the receiver operating characteristic curve across *P*-value cutoffs was used, comparing the area under the curve (AUC) for each. An AUC of less than 0.5 demonstrated that for most taxa in this simulated scenario cMWU and cTT performed no better than guessing.

Next, the classification accuracy was compared of the three procedures designed to utilize spike-in quantification data, sMWU, sTT, and SpikeAbund. All three performed substantially better than random guessing in abundant and moderately abundant taxa (MCC > 0). For rare and very rare taxa, the accuracy of the two naïve approaches, sMWU and sTT, was moderately lower than SpikeAbund, largely driven by a much worse FDR. Compared to the other two, the FDR for SpikeAbund was better controlled in rare taxa, although sensitivity was also lower, correctly reflecting the reduced information content with regards to these taxa.

To examine our ability to estimate effect sizes from spike-in quantification experiments, the known, true effect parameters are compared to point estimates obtained from the simulated data (see Figure 5.2). For the most abundant taxa, the mean squared error of the estimate was substantially lower for both methods that included spike information, and particularly for SpikeAbund. While the MSE of point estimates from SpikeAbund were much larger for rare taxa, this is apparently due to inherent uncertainty about the true abundance. Since rare taxa may vary in relative abundance over many orders of magnitude, while still remaining below the limit of detection, large relative effects of covariates on abundance cannot be ruled out. Despite the inflated MSE, the inaccuracy of these results is appropriately reflected in the width of the posterior distribution. Across the 50 taxa in 1,000 simulations, true parameter values were within the 95% highest probability density interval for 94.6% of them, almost exactly the expected fraction. This suggests that credible intervals obtained using SpikeAbund reflect a reasonable state of belief given the data.

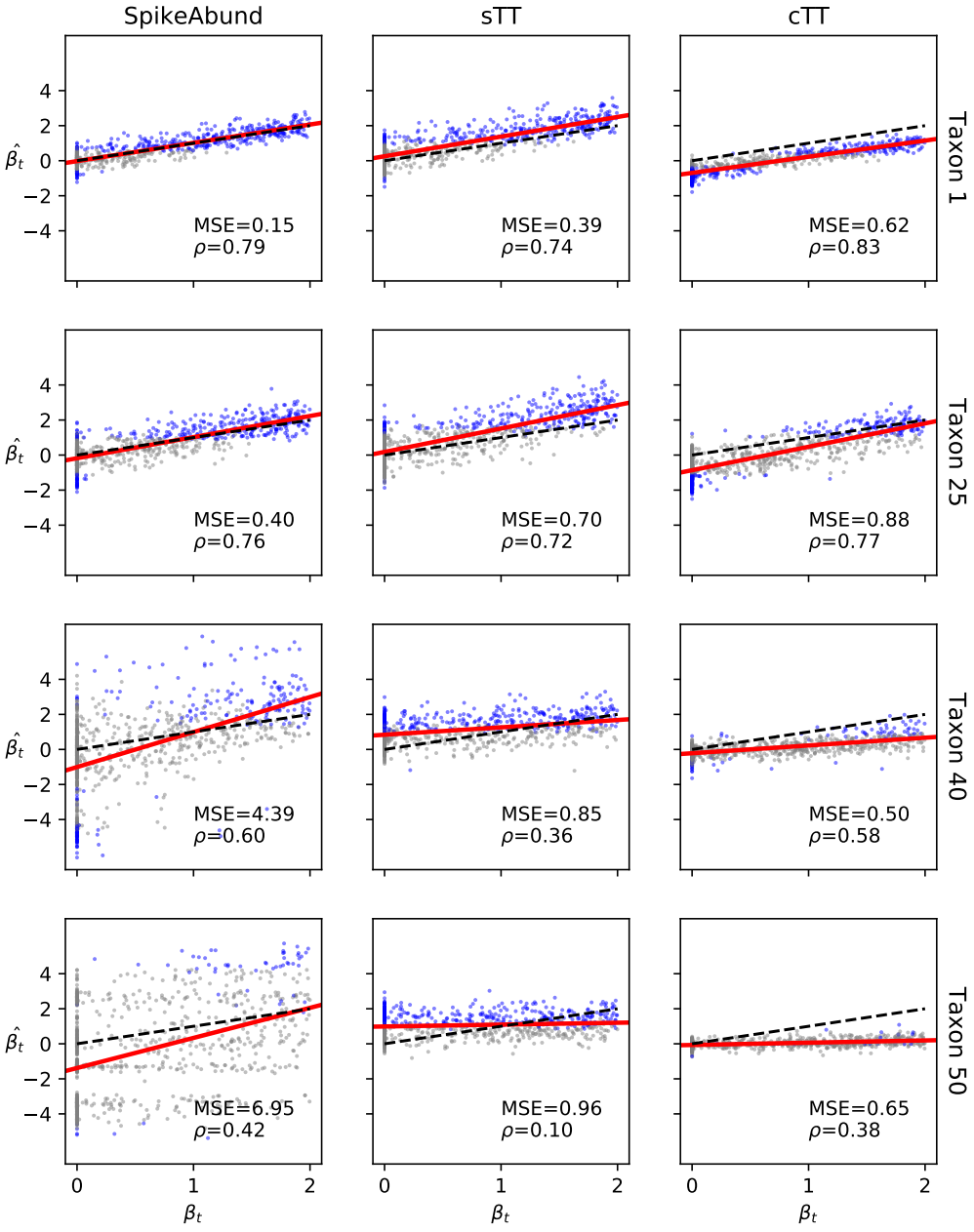
SpikeAbund
sTT
cTT
Taxon 1
Taxon 25
Taxon 40
Taxon 50
$\hat{\beta}_t$
β_t
MSE=0.15
ρ=0.79
MSE=0.39
ρ=0.74
MSE=0.62
ρ=0.83
MSE=0.40
ρ=0.76
MSE=0.70
ρ=0.72
MSE=0.88
ρ=0.77
MSE=4.39
ρ=0.60
MSE=0.85
ρ=0.36
MSE=0.50
ρ=0.58
MSE=6.95
ρ=0.42
MSE=0.96
ρ=0.10
MSE=0.65
ρ=0.38

Figure 5.2: Parameter estimates made by SpikeAbund vs. two naïve methods. True parameter values are compared to estimates obtained using three analysis methods (columns) for four taxa (rows) representing high abundance (Taxon 1), moderate abundance (Taxon 25), low abundance (Taxon 40) and very low abundance (Taxon 50). In each panel, points corresponds to each of 1,000 simulations. Simulations in which that taxon was identified as significantly differentially abundant are indicated (blue points). The one-to-one line is shown (black, dashed), as well as a regression between the true and estimated parameters (red line).

5.5 Identification of bacteria affected by acarbose treatment in mice

To demonstrate the utility of this method in the analysis of real data, SpikeAbund was applied to 16S rRNA gene libraries from 143 fecal samples collected in a previous study of acarbose in mice (see Chapter 2), in order to identify taxa with density affected by treatment with the drug. In that data, 105 operation taxonomic units (OTUs) were "common": with mean relative read abundance greater than 0.1% and found in at least 5% of samples. The remaining reads were pooled and treated as a single taxon. The set of taxa flagged by this approach are compared to results from MWU tests on the relative abundance of each OTU.

Of the 105 common OTUs in this data, the effect of acarbose on 60 of these was confidently assessed, defined as a posterior probability of a type-S error of less than 0.05. Of these, 16 were positively and 44 were negatively affected, suggesting that acarbose treatment was deleterious to more taxa than it was beneficial. Of particular note, SpikeAbund confirmed that the absolute abundance of OTU-1 increased with acarbose treatment, and not only relative abundance as reported in Chapter 2. Of these 60, the naïve test on relative abundance reaffirmed the classification for all but 6. Interestingly, for all of these the mean posterior value of the effect parameter was greater than zero, suggesting evidence for a positive effect of acarbose. Conversely, the additional 8 OTUs flagged only by the test on relative abundance, all showed trends towards a negative effect. This may reflect the obfuscating impact of a large increase in OTU-1 on the ability to detect changes in other taxa when only considering relative abundance.

One case of particular interest is OTU-49, flagged by SpikeAbund as negatively affected, where the sign of the inferred effect on absolute abundance contradicts the apparent positive effect on relative abundance. However, a statistical test of the latter was not significant and the estimated magnitude was small. The general concordance between inferences on absolute abundance, using SpikeAbund, and relative abundance suggest that total community size in this experiment was not substantially affected by acarbose treatment. While there is not evidence that relative abundance would have been directly misleading in this experimental system, use of spike-in quantification enables us to account for this possibility in our interpretations.

5.6 Model criticism and improvement

The correctness of conclusions drawn from this procedure are dependent on the ability of the model itself to capture relevant population dynamics. For this reason, model criticism in light of real data constitutes an important component of the scientific process.

Towards this end, we compared the marginalized posterior distribution of d for each sample i (equal to $\sum_{j=1}^{J} y_{ij}$) to point estimates of the parameter ($\hat{d}$) obtained using Equation 5.1. This parameter is a proxy for total 16S rRNA gene density. For one sample in particular, labeled JL0836, these deviated substantially (see Figure 5.3 panel A), with the posterior median nearly two orders of magnitude larger than $\hat{d}$, and nearly one order of magnitude larger than for the sample with the next highest value. This unreasonably inflated posterior, along with the generally poor concordance between point estimates of d and the posterior distribution of the parameter for other samples, suggests that the probabilistic model developed here does not perfectly capture the true data generating process being studied.

Further examination of sample JL0836 found that the total number of common OTUs observed, not counting the spike, was only 12, making it by far the least taxonomically diverse of the 143 samples. A majority, 58%, of sequences in this library were clustered into OTU-34, classified as a member of the genus *Klebsiella* which includes known opportunistic pathogens. Given this observation, the dominance of OTU-34 in the fecal community of JL0836 may indicate an active infection that also resulted in much lower densities of non-pathogenic bacteria in that sample. We propose that the anomalous posterior distribution of d for this sample reflects a failure of the model to account for broad correlations in the latent abundances of taxa, y_j, within samples. When these correlations arise, for example due to dysbiosis brought on by a pathogen, the model does not reflect our *a priori* understanding of the system. If this is indeed the problem here, amending the model to account for global correlations among taxa will better fit the data, and will align estimates of d with our intuition.

The model was updated by adding a normally distributed, sample-specific term, γ_i, to the log density of all taxa:

$$
\begin{aligned}
Z_i &\sim \mathrm{DM}(\alpha\Pi_i, m) \\
\Pi_i &= \left[\frac{y_{i1}}{s_i + d_i}, \ldots, \frac{y_{iJ}}{s_i + d_i}, s_i\right] \\
d_i &= \sum_{j=1}^{J} y_{ij} \\
\log(y_{ij}) &\sim \mathrm{Norm}(\mu_{ij}, \sigma_j) \\
\mu_{ij} &= \beta_{0j} + \sum_{k=1}^{K} x_{ik}\beta_{kj} + \gamma_i \\
\gamma_i &\sim \mathrm{Norm}(0, \psi)
\end{aligned}
$$

where the parameter ψ describes the strength of the global correlation in taxon abundance. We refer to this adjusted version as Model 2, and the original as Model 1. Conditioning on Model 2, the posterior distribution of d for JL0836 is found to be consistent with the point estimate, $\hat{d}$. In fact, this greater concordance is shared by all samples (see Figure 5.3 panel B). Comparison of the two models using the widely applicable information criterion (WAIC) [31] shows overwhelming evidence that Model 2 better describes the data ($\Delta\mathrm{WAIC} = 13091.7$).

Unsurprisingly, the posterior distributions of the linear model parameters, β, are different when conditioning on Model 2 (see Figure 5.4). While inferences about the effects of acarbose on bacterial abundance are therefore potentially sensitive to the choice of model, 54 of the 60 taxa originally flagged as affected by acarbose, were also identified when using Model 2.

An iteration of criticism and refinement here results in a new model that appears to better describe the abundances of bacteria in the microbial community. While additional study is needed to asses in other systems the relevance of the proposed "global" correlation among taxa, spike-in quantification paired with model based inference makes further exploration feasible.

5.7 Conclusions

We have introduced a novel statistical procedure for the analysis of data resulting from spike-in quantification experiments, and have found that under one plausible scenario it performs better than naïve method operating on either relative abundance or absolute abundance point estimates. On real data, this method is able to identify

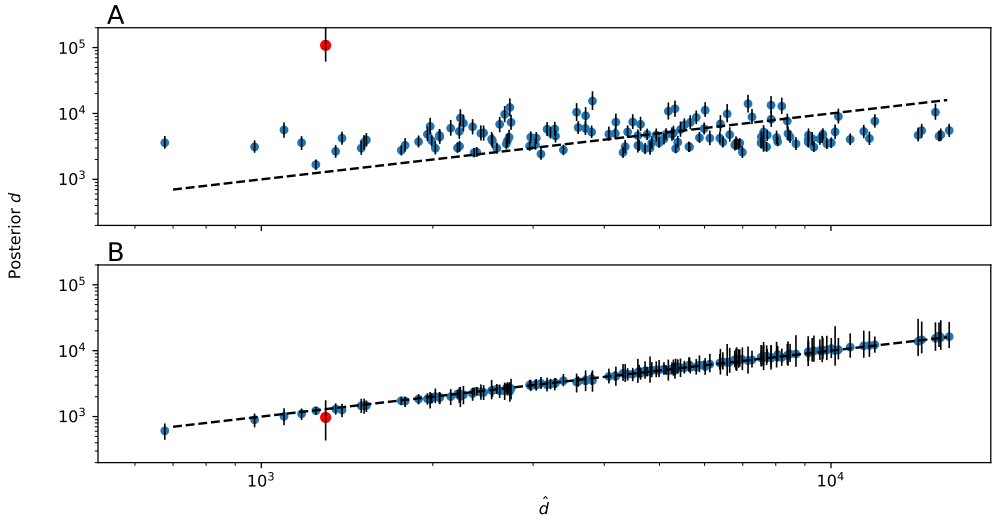

Figure 5.3: Posterior distributions versus point estimates using two models. Point estimates of d derived from Equation 5.1 are plotted against the marginalized posterior distribution. Posteriors are conditioned on either (**A**) Model 1, or (**B**) Model 2 that also includes a term, γ, correlating taxon abundances across samples. Both the posterior median (circles) and the 95% credible interval (black bars) are shown. The 1-to-1 line is plotted (dashed line) and the anomalous sample JL0836 is highlighted (red circle) in both panels.

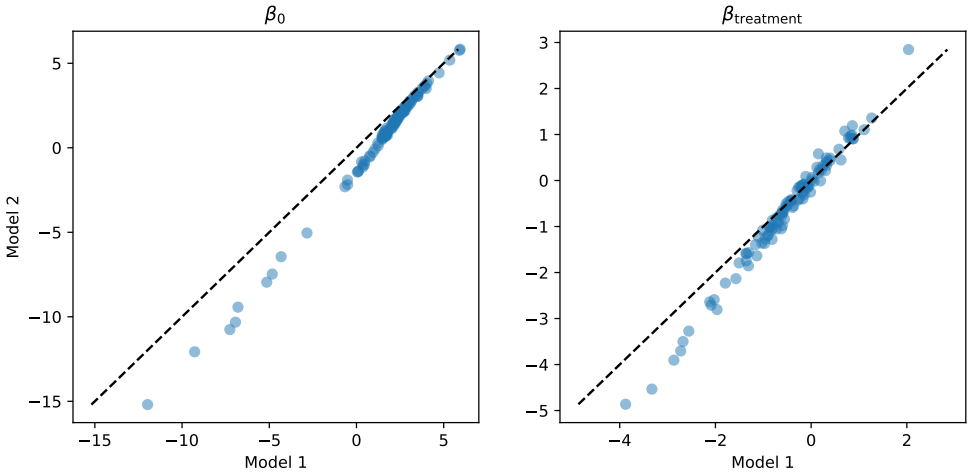

Figure 5.4: Comparison of linear model parameter estimates under two models. Points correspond with individual taxa and the posterior medians are shown for, β_0, the log baseline abundance of each taxon, and $\beta_{\text{treatment}}$, the effect of acarbose on the abundance of each taxon. The relationship between estimates conditioned on each model is shown. Model 1 and Model 2 differ by the inclusion in the latter of a parameter, γ, correlating taxon abundances across samples. The 1-to-1 line (dashed) is shown.

taxa affected by experimental manipulations. By harnessing a model-based approach, the inherent assumptions are made explicit, presenting a platform for iterative refinement and criticism of theory with data. The flexibility of the approach fosters further refinement of the basic community abundance model. For instance, inter-taxon correlations are a simple addition, and hierarchical effects on taxa, such as phylogenetically constrained effects, are also possible. Likewise, expansion of this procedure to other experimental designs such as random blocks and repeated measures is a straightforward extension.

While other methods exist to measure bacterial absolute abundance, spike-in quantification is simple and widely applicable. Combining these data with the model-based inference described here will generate novel insights into microbial community dynamics inaccessible to current approaches.

5.8 Methods

5.8.1 Simulation of realistic data

To explore statistical approaches for the identification of differential abundance, communities, were numerically simulated based on a hypothesized dietary polysaccharide supplementation experiment. For each simulation, 50 taxa were generated, 25 of these were randomly designated as responders and had their treatment response parameter, β_t sampled from a Uniform(0, 2) distribution. The remaining 25 taxa were designated as non-responders and had their response parameter set to 0. Taxa were randomly permuted and assigned a rank from 1 to 50. The baseline log abundance, β_0, for taxon rank i was set to

$$\beta_0 = (i-1)\frac{\log(a) - \log(b)}{k-1} + \log(a)$$

where a is the abundance of taxon rank 1 ($a = 10$) and b is the abundance of taxon rank 50 ($b = 0.001$). This formulation is designed so that baseline abundances quickly decrease in lower ranks relative to the dominant taxa.

For each experiment, 15 control and 15 treated replicate communities were sampled. The log absolute abundance for all taxa in each replicate was sampled from a normal distribution with $\sigma = 0.5$ and $\mu = \beta_0 + \beta_t X$ where X is set to 1 for treated replicates and 0 otherwise. Spike abundance for each replicate was set to 1. Read counts were simulated for each replicate from a Dirichlet-multinomial distribution parameterized with the relative abundance of each taxon (including the spike), a concentration parameter (α) of 100, and a sample size of 5000 reads.

5.8.2 cMWU, cTT, sMWU, sTT implementation

Four alternative statistical procedures were chosen to represent likely approaches to data analysis by relatively naïve practitioners. Two of the procedures operate directly on compositional data, without considering counts of the spike sequence. One of these (cMWU) uses a non-parametric test, the Mann-Whitney U, frequently applied to the analysis of non-normal data, including microbial relative abundance [e.g. 32]. The other (cTT) takes a parametric approach, utilizing a standard, two-sample t-test, and log-transforming relative abundance to improve distributional characteristics. Because the logarithm of 0 is undefined, a pseudocount of 1 was added to each cell

of the count table before calculating relative abundance. Two additional procedures were chosen that leverage point estimates of absolute abundance (also calculated with pseudocounts) to go beyond composition. These also utilized a Mann-Whitney U test (sMWU) and a t-test on log-transformed values (sTT). Estimates of effect size associated with cTT and sTT were calculated from relative and absolute abundance data as the difference in means of log values.

5.8.3 Acarbose experiment data and analysis

Experimental data was used from the study described in Chapter 2 on acarbose treatment in mice. 16S rRNA gene libraries are available from the SRA, and scripts to automatically process these data are made available online [33]. Briefly, in that study samples (n = 143) were spiked with whole-cell cultures of *Sphingopyxis alaskensis* and normalized to the wet weight of the sample. OTUs were clustered at a 97% identity threshold.

SpikeAbund was used to fit a linear model to these data with terms for treatment, study site, sex of the mouse, and the interaction between treatment and sex (in R-style notation: abundance $\sim$ treatment $\times$ sex + site$'$). Default priors ($\sigma = 10$) were placed on all 6 model parameters for all OTUs. The marginal posterior distributions on $\beta_{\text{treatment}}$ were interpreted for each OTU. The mean of the posterior was used as a point estimate and the fraction of MCMC samples with different sign than the mean was used by analogy to a traditional *P*-value. Fractions less than 0.05 were flagged as significantly different between control and treatment.

BIBLIOGRAPHY

[1] Tringe, S.G., Hugenholtz, P.: A renaissance for the pioneering 16S rRNA gene. Current Opinion in Microbiology **11**(5), 442–446 (2008). doi:10.1016/j.mib.2008.09.011

[2] Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., Gordon, J.I.: The Human Microbiome Project. Nature **449**(7164), 804–810 (2007). doi:10.1038/nature06244. arXiv:1011.1669v3

[3] Turnbaugh, P.J., Ley, R.E., Mahowald, M.A.: An obesity-associated gut microbiome with increased capacity for energy harvest. Nature **444** (2006)

[4] Rubin, T.a., Gessert, C.E., Aas, J., Bakken, J.S.: Fecal microbiome transplantation for recurrent Clostridium difficile infection: Report on a case series. Anaerobe (November), 1–5 (2012). doi:10.1016/j.anaerobe.2012.11.004

[5] Tsagris, M.: Regression analysis with compositional data containing zero values (2008), 1–12 (2015). arXiv:1508.01913v1

[6] Mandal, S., Treuren, W.V., White, R.A., Eggesbø, M., Knight, R., Peddada, S.D.: Analysis of composition of microbiomes: a novel method for studying microbial composition. Microbial Ecology in Health and Disease **26**, 27663 (2015). doi:10.3402/mehd.v26.27663

[7] Lovell, D., Pawlowsky-Glahn, V., Egozcue, J.J., Marguerat, S., Bähler, J.: Proportionality: A Valid Alternative to Correlation for Relative Data. PLOS Computational Biology **11**(3), 1004075 (2015). doi:10.1371/journal.pcbi.1004075

[8] Kurtz, Z.D., Müller, C.L., Miraldi, E.R., Littman, D.R., Blaser, M.J., Bonneau, R.A.: Sparse and Compositionally Robust Inference of Microbial Ecological Networks. PLOS Computational Biology **11**(5), 1004226 (2015). doi:10.1371/journal.pcbi.1004226

[9] McMurdie, P.J., Holmes, S.: Waste not, want not: why rarefying microbiome data is inadmissible. PLoS computational biology **10**(4), 1003531 (2014). doi:10.1371/journal.pcbi.1003531

[10] O'Brien, J.D., Record, N.: The power and pitfalls of Dirichlet-multinomial mixture models for ecological count data. bioRxiv, 045468 (2016). doi:10.1101/045468

[11] Biswas, S., Mcdonald, M., Lundberg, D.S., Dangl, J.L., Jojic, V.: Learning Microbial Interaction Networks from Metagenomic Count Data. Journal of Computational Biology **23**(6), 526–535 (2016). doi:10.1089/cmb.2016.0061

[12] Morton, J.T., Sanders, J., Quinn, R.A., Mcdonald, D., Gonzalez, A., Vázquez-baeza, Y., Navas-molina, J.A.: Balance trees reveal microbial niche differentiation **2**(1), 1–11 (2017). doi:10.1128/mSystems.00162-16

[13] Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., Xia, L.C., Xu, Z.Z., Ursell, L., Alm, E.J., Birmingham, A., Cram, J.A., Fuhrman, J.A., Raes, J., Sun, F., Zhou, J., Knight, R.: Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. ISME Journal **10**(7), 1669–1681 (2016). doi:10.1038/ismej.2015.235

[14] Stämmler, F., Gläsner, J., Hiergeist, A., Holler, E., Weber, D., Oefner, P.J., Gessner, A., Spang, R.: Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. Microbiome, 1–13 (2016). doi:10.1186/s40168-016-0175-0

[15] Smets, W., Leff, J.W., Bradford, M.A., McCulley, R.L., Lebeer, S., Fierer, N.: A method for simultaneous measurement of soil bacterial abundances and community composition via 16S rRNA gene sequencing. Soil Biology and Biochemistry **96**, 145–151 (2016). doi:10.1016/j.soilbio.2016.02.003

[16] Morel, J.G., Nagaraj, N.K.: A finite mixture distribution for modelling multinomial extra variation. Biometrika **80**(2), 363–371 (1993). doi:10.1093/biomet/80.2.363

[17] Costea, P.I., Hildebrand, F., Arumugam, M., Bäckhed, F., Blaser, M.J., Bushman, F.D., de Vos, W.M., Ehrlich, S.D., Fraser, C.M., Hattori, M., Huttenhower, C., Jeffery, I.B., Knights, D., Lewis, J.D., Ley, R.E., Ochman, H., O'Toole, P.W., Quince, C., Relman, D.A., Shanahan, F., Sunagawa, S., Wang, J., Weinstock, G.M., Wu, G.D., Zeller, G., Zhao, L., Raes, J., Knight, R., Bork, P.: Enterotypes in the landscape of gut microbial community composition. Nature Microbiology **Accepted**(January) (2017). doi:10.1038/s41564-017-0072-8

[18] Holmes, I., Harris, K., Quince, C.: Dirichlet multinomial mixtures: Generative models for microbial metagenomics. PLoS ONE **7**(2) (2012). doi:10.1371/journal.pone.0030126

[19] Mattiello, F., Verbist, B., Faust, K., Raes, J., Shannon, W.D., Bijnens, L., Thas, O.: A web application for sample size and power calculation in case-control microbiome studies. Bioinformatics **32**(February), 099 (2016). doi:10.1093/bioinformatics/btw099

[20] Mimno, D., McCallum, A.: Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression (2012). doi:10.1.1.140.6925. 1206.3278

[21] Bouguila, N.: Clustering of count data using generalized dirichlet multinomial distributions **20**(4), 462–474 (2008)

[22] Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., Sahl, J.W., Stres, B., Thallinger, G.G., Van Horn, D.J., Weber, C.F.: Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology **75**(23), 7537–7541 (2009). doi:10.1128/AEM.01541-09

[23] Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.a., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.a., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.a., Widmann, J., Yatsunenko, T., Zaneveld, J., Knight, R.: QIIME allows analysis of high-throughput community sequencing data. Nature Methods **7**(5), 335–336 (2010). doi:10.1038/nmeth.f.303. NIHMS150003

[24] Salvatier, J., Wiecki, T., Fonnesbeck, C.: Probabilistic Programming in Python using PyMC. Arxiv, 1–24 (2015). doi:10.7717/peerj-cs.55. 1507.08050

[25] patsy - Describing statistical models in Python. https://patsy.readthedocs.io/en/latest/ Accessed 2018-11-08

[26] Hoffman, M.D., Gelman, A.: The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo **15**, 1593–1623 (2011). doi:10.1190/1.3627885. 1111.4246

[27] Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., Gelman, A.: Visualization in Bayesian workflow **2** (2017). doi:10.1016/S0013-7006(08)73280-9. 1709.01449

[28] Gelman, A., Rubin, D.B.: Inference from Iterative Simulation Using Multiple Sequences. Statistical Science **7**(4), 457–472 (1992). doi:10.1214/ss/1177011136. arXiv:1011.1669v3

[29] Gelman, A., Tuerlinckx, F.: Type S error rates for classical and Bayesian single and multiple comparison procedures. Computational Statistics (2000)

[30] Gelman, A., Carlin, J.: Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. Perspectives on Psychological Science **9**(6), 641–651 (2014). doi:10.1177/1745691614551642

[31] Watanabe, S.: Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory **11**, 3571–3594 (2010)

[32] Choo, J.M., Leong, L.E.X., Rogers, G.B.: Sample storage conditions significantly influence faecal microbiome profiles. Scientific Reports **5**, 1–10 (2015). doi:10.1038/srep16350. 1310.0424

[33] Smith, B.J.: Code and Metadata to Reproduce: Changes in the gut microbiota and fermentation products associated with enhanced longevity in acarbose-treated mice. (2018). doi:10.5281/zenodo.1229203. https://github.com/bsmith89/smith2018paper/releases/tag/v0.1 Accessed 2018-04-25

CHAPTER 6

Summary and Conclusions

The four chapters of this dissertation are loosely joined by a shared theme of deepening understanding of complex microbial communities and their impacts using varied data and improved tools. I was fortunate to have the opportunity, three years ago, to take part in the Interventions Testing Program, giving me access to a wealth of data not often available in microbiome studies. This collaboration forms the biological and experimental core of the work presented here, and serves as a platform for both my improved understanding of host-associated microbial communities and the tools we use to explore them. This has resulted in a new perspective on the potential role of bacterial fermentation products in mouse longevity, the impact of acarbose on gut microbial communities, and the ecological niche of *Muribaculaceae*. Simultaneously, it has driven development of tools and techniques for integrating disparate data types, inferring physiological properties of uncultured bacteria, and overcoming the limitations of marker gene surveys. This top-down approach to biology has become centrally important as the field struggles to match the increasing rate of data collection with improvements in mechanistic understanding. Achieving this goal will require not just new data, new models, and new tools, but also a new mindset in order to leverage these for improved prediction, explanation, and manipulation.

The four studies in my dissertation contribute directly towards this end.

In Chapter 2, I describe the response of microbial communities in the guts of mice to the longevity enhancing drug acarbose. This study combined targeted measurements of fecal metabolites, 16S rRNA gene surveys, and mouse survival data. Using these, I was able to explore predictions of the exciting hypothesis that the observed life-extending effects of acarbose in mice result from the increased production of short-chain fatty acids by gut bacteria. I demonstrated that community composition was changed by acarbose, that this was linked with changes in the concentrations of fecal metabolites, and that these metabolites were associated with mouse lifespan.

I also found bacterial taxa that responded dramatically to acarbose treatment. This chapter exemplifies the power of integrating multiple data types and comprehensive exploration of microbial communities.

In Chapter 3 I compared the genomes of the two bacterial taxa that responded to acarbose treatment, to five "non-responders", all in the largely uncultured family *Muribaculaceae*. This was achieved through careful reconstruction of metagenomic reads, resulting in high-quality inferred genomes. Unlikely previous culture-free studies of the clade, which were done in the absence of any physiological information, ours leveraged the observed difference in acarbose response to explore hypotheses about the ecological niche occupied by members of this family. I demonstrated that responders have genomic features consistent with starch utilization, while non-responders generally do not. I was also able to identify two distinct genomic variants, potentially explaining site specific dynamics. This chapter harnesses modern bioinformatic approaches to better understand the physiology of bacteria without cultured representatives.

In Chapter 4 I present a detailed perspective on a new approach to quantification of bacterial communities. Microbial ecology, particularly studies of host-associated microbiomes, is severely limited by the absence of absolute abundance information. While sophisticated analyses can avoid the major pitfalls of compositional data, interpretation of relative abundance is still constrained. I presents evidence that spike-in quantification is robust and provides valuable abundance information, and I discuss features that make it an attractive option relative to other methods, such as qPCR. I also suggest a set of best practices based on my experience with the approach. New experimental tools will be an important part of microbial ecology in the coming decade. This chapter presents a case for using one such tool, and attempts to make it accessible to others in the field.

In Chapter 5 I expand on my treatment of this approach, presenting a statistical model for spike-in quantification experiments and then applying it for inference of community dynamics. I introduce a software tool for others to apply my method to their own experiments, and demonstrate its value in simulated and real data. I also describe future extensions to this model that will leverage the newly accessible absolute abundance information to explore complex community features. Particularly in host-associated studies, *ex vivo* manipulations are well complemented by techniques that can harness *in vivo* observations for improved understanding. Model-based approaches to data analysis have the potential to expand our understanding of microbial communities in cases where direct measurements are challenging or impossible.

Combined, these chapters explore deeply the features of the fecal microbiota in acarbose treated mice and controls. I introduce a microbiome perspective to an experimental model primarily studied in terms of host physiology. In this system, I have found and explored a wealth of microbiological and ecological phenomena, a portion of which have the potential to explain features of host health. While more extensive study will be necessary to fully understand the role of gut microbes in acarbose response and longevity, my dissertation has provided a foundation for this future work.

In the more than seven years of my PhD, I have developed an appreciation for my own strengths and interests, as well as my limitations. My evolution as a scientist is reflected in the chapters of this dissertation, but perhaps even more so in the myriad past and present projects that have not resulted in publications. Given my respect for bottom up approaches and experimentation, it has been at times challenging to negotiate a strong personal affinity for computational and top-down approaches. But, as I move on to the next stages in my career, I am eager to focus on my strengths. I believe my opportunities for impact will be in extracting understanding from the large, messy data that results from modern experimental approaches.

My postdoctoral work will continue the focus on host-associated microbial communities, this time studying the ecological and health impacts of fecal microbiome transplants in patients with ulcerative colitis. I will be further developing computational approaches for recovering genomes from metagenomes, with the goal of understanding variation in the function and persistence of bacteria in the gut. I feel fortunate to be working in a field with not only the potential to improve human health, but where I can straddle the space between microbial ecologist, bioinformatician, statistician, and data scientist. My doctoral work has provided me with invaluable experience and perspective for my coming career.

APPENDIX A

Taxonomic analysis of two dominant OTUs in acarbose treated mice

The most conspicuous difference in the gut microbiota induced by ACA was a site-specific effect in two populations of bacteria both classified as members of family *Muribaculaceae* (Figure 2.3). OTUs were classified based on an approximately 240 bp fragment of the 16S rRNA gene in the V4 hypervariable region. Using this fragment, we applied several lines of evidence to confirm that OTU-1 and OTU-4 are both members of the *Muribaculaceae* and that they are genetically distinct from cultured relatives. Classification of sequences using the method of Wang *et al.* [1] and the SILVA non-redundant database as a reference [2], identified both OTU-1 and OTU-4 as members of the family with 100% bootstrap support. While use of the RDP training set [3] Version 14 instead assigned these sequences to the family Porphyromonadaceae this is presumably because the *Muribaculaceae* are not recognized as a taxon in the RDP (previously reported by [4]), nor are alternative names for the clade ("S24-7" or "*Homeothermaceae*"). A follow-up phylogenetic analysis of representative amplicon sequences from two dominant OTUs was carried out using approximate maximum likelihood estimation implemented in the FastTree software (version 2.1.8 [5]) using the generalized time reversible model with twenty discrete rate categories (`-gtr -gamma` options). Approximate maximum likelihood phylogenetic estimation, using a selection of type strains in the order *Bacteroidales*, places OTU-1 and OTU-4 in a clade with representatives of the *Muribaculaceae* with >95% support for the topology of that node (see Figure A.1). While such a short sequence fragment is unlikely to perfectly recapitulate phylogeny—indeed, tree topology was generally weakly supported and was sensitive to both the choice of reference sequences and the evolutionary model used—we are nonetheless satisfied with the evidence for assignment of both OTU sequences to this clade; besides exceptions in the *Porphyromonadaceae*, *Marinilabiliaceae*, and *Bacteroides*, our phylogenetic reconstruction largely matches

a recently proposed taxonomy of the *Bacteroidales* [6].

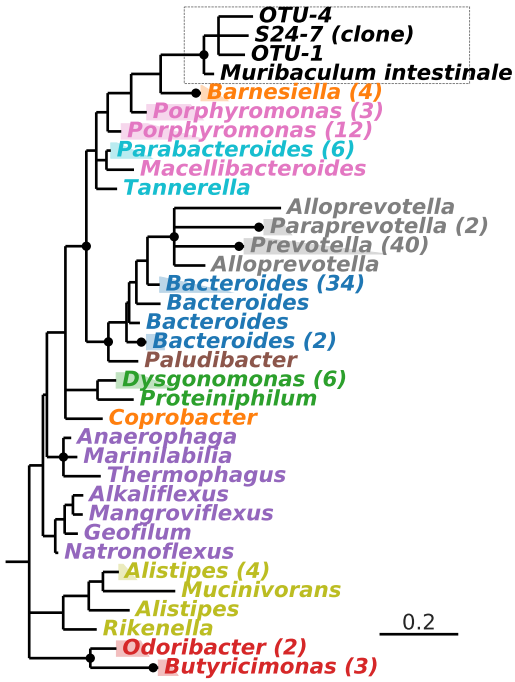

Figure A.1: Phylogenetic characterization of OTU-1 and OTU-4. Estimated phylogeny based on approximately 240 bp of the 16S rRNA gene V4 hypervariable region and more than 130 type strain reference sequences spanning the diversity of order *Bacteroidales*. Branch lengths are in units of expected substitutions per site. The tree is rooted by a *Flavobacteriales* out-group (not shown). Reference taxa are labeled with genus designations according to the SILVA database. When multiple representatives from the same genus have been folded together, the number of sequences is reported in parentheses. Nodes with Shimodaira-Hasegawa local support over 95% are indicated with black circles and nodes with support less than 70% have been collapsed to polytomies. The dashed box encloses taxa inferred to be within the *Muribaculaceae*. The taxon labeled 'S24-7 (clone)' (GenBank: AJ400263.1) is the environmental sequence by which the clade was originally identified, and by which it was historically named [7], while *Muribaculum intestinalis* (GenBank: KR364784.1) is the first cultured representative [8]. Label colors indicate a recently proposed family membership of each reference: *Prevotellaceae* (gray), *Barnesiellaceae* (orange), *Porphyromonadaceae* (pink), *Dysgonomonadaceae* (green), *Bacteroidaceae* (blue), *Tannerellaceae* (light blue), *Marnilabilaceae* (purple), *Marinifilaceae* (red), *Paludibacteraceae* (brown), and *Rikenellaceae* (yellow) [6].

OTU-1 and OTU-4 represent uncultured genera. Over the analyzed sequence they have 89% and 92% identity, respectively, to *Muribaculum intestinalis* strain YL27, the first cultured representative of the *Muribaculaceae* [8]. A BLAST search against the

NCBI non-redundant nucleotide collection did not find higher sequence similarity to any other cultured bacteria. Representative sequences for OTU-1 and OTU-4 share nucleotides at only 22 out of 244 positions (91%).

BIBLIOGRAPHY

[1] Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R.: Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology **73**(16), 5261–5267 (2007). doi:10.1128/AEM.00062-07. Wang, Qiong, 2007, Naive

[2] Yilmaz, P., Parfrey, L.W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., Schweer, T., Peplies, J., Ludwig, W., Glöckner, F.O.: The SILVA and "all-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Research **42**(D1), 643–648 (2014). doi:10.1093/nar/gkt1209

[3] Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R., Tiedje, J.M.: Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Research **42**(Database issue), 633–642 (2014). doi:10.1093/nar/gkt1244

[4] Clavel, T., Lagkouvardos, I., Blaut, M., Stecher, B.: The mouse gut microbiome revisited: From complex diversity to model ecosystems. International Journal of Medical Microbiology **306**(5), 316–327 (2016). doi:10.1016/j.ijmm.2016.03.002

[5] Price, M.N., Dehal, P.S., Arkin, A.P.: FastTree 2 - approximately maximum - likelihood trees for large alignments. PloS one **5**(3), 9490 (2010). doi:10.1371/journal.pone.0009490

[6] Ormerod, K.L., Wood, D.L.A., Lachner, N., Gellatly, S.L., Daly, J.N., Parsons, J.D., Dal'Molin, C.G.O., Palfreyman, R.W., Nielsen, L.K., Cooper, M.A., Morrison, M., Hansbro, P.M., Hugenholtz, P.: Genomic characterization of the uncultured Bacteroidales family S24-7 inhabiting the guts of homeothermic animals. Microbiome **4**(1), 36 (2016). doi:10.1186/s40168-016-0181-2

[7] Salzman, N.H., de Jong, H., Paterson, Y., Harmsen, H.J.M., Welling, G.W., Bos, N.A.: Analysis of 16S libraries of mouse gastrointestinal microflora reveals a large new group of mouse intestinal bacteria. Microbiology **148**(11), 3651–3660 (2002). doi:10.1099/00221287-148-11-3651

[8] Lagkouvardos, I., Pukall, R., Abt, B., Foesel, B.U., Meier-Kolthoff, J.P., Kumar, N., Bresciani, A., Martínez, I., Just, S., Ziegler, C., Brugiroux, S., Garzetti, D., Wenning, M., Bui, T.P.N., Wang, J., Hugenholtz, F., Plugge, C.M., Peterson, D.A., Hornef, M.W., Baines, J.F., Smidt, H., Walter, J., Kristiansen, K., Nielsen, H.B., Haller, D., Overmann, J., Stecher, B., Clavel, T.: The Mouse Intestinal

Bacterial Collection (miBC) provides host-specific insight into cultured diversity and functional potential of the gut microbiota. Nature Microbiology **1**(August), 16131 (2016). doi:10.1038/nmicrobiol.2016.131

APPENDIX B

Expanded survival analysis of ITP mice

Proportional hazards regression was used in this study to demonstrate an association between SCFA concentrations in feces and the longevity of mice. Given the limited number of samples for which matched chemical and survival data are available, statistical testing of associations with SCFAs were carried out in the pooled dataset. Therefore, to account for known effects of treatment, sex, and site, the primary null model used in this study includes terms for all main, two, and three-way interactions of the design covariates: treatment, sex, and study site. A priori, a number of these terms were expected to be non-zero based on ITP findings from previous cohort years [1, 2]. Indeed, when survival data from all of control and ACA treated mice in this cohort were analyzed together, effects were detected that recapitulated these expectations, including: increased longevity of females, increased longevity with ACA treatment, and increased longevity of control males at UM (see Table B.1).

Table B.1: Fitted coefficients for experimental covariates in the full ITP cohort

Term	log(HR)	Std. Error	P
ACA	-0.530	0.170	0.002
Female	-0.266	0.143	0.063
TJL	-0.024	0.144	0.865
UM	-0.598	0.155	0.000
ACA:Female	0.230	0.245	0.349
ACA:TJL	-0.225	0.242	0.353
ACA:UM	0.003	0.254	0.992
Female:TJL	-0.017	0.204	0.934
Female:UM	0.580	0.213	0.006
ACA:Female:TJL	0.351	0.349	0.313
ACA:Female:UM	-0.054	0.361	0.881

Interestingly, some—though not all—of these effects were still evident when analyzing the much smaller data set of mice from which we collected fecal samples.

Table B.2: Survival effect estimates for experimental covariates

Term	log(HR)	Std. Error	P
ACA	-0.770	0.426	0.071
female	-0.239	0.421	0.570
UM	-0.931	0.439	0.034
ACA:female	0.398	0.589	0.498
ACA:UM	0.146	0.608	0.810
female:UM	0.807	0.597	0.176
ACA:female:UM	0.187	0.845	0.825

This increased our confidence that, despite the age of the mice at the time of entry and the limited sample size, the associations with SCFA concentrations reflect patterns that could be seen in the full cohort.

As described in the main results, adding terms for the concentrations of three SCFAs improved the fit of the model (see Table B.3).

Table B.3: Survival effect estimates for experimental covariates and SCFAs

Term	log(HR)	Std. Error	P
propionate	-0.292	0.116	0.012
butyrate	-0.119	0.055	0.030
acetate	0.062	0.030	0.042
ACA	-0.124	0.488	0.799
female	-0.476	0.433	0.272
UM	-1.232	0.484	0.011
ACA:female	0.095	0.597	0.874
ACA:UM	0.241	0.654	0.713
female:UM	0.913	0.614	0.137
ACA:female:UM	0.111	0.850	0.896

Interestingly, the positive association betwen ACA and longevity is distinctly weakened when SCFAs are included (estimated coefficient went from -0.770 without SCFAs to -0.124 with). Although we do not carry out a formal path analysis, this is consistent with the causal effects of ACA on longevity being mediated by SCFA concentrations.

Survival analysis is potentially sensitivity to deviations from the assumptions of the Cox family of models [3]. We therefore tested proportionality and linearity assumptions relevant to our main findings. The Cox proportional hazards model assumes that the hazard associated with each covariate is proportional across the full set of ages at which mice are being tracked—e.g. that the proportional decrease in the

risk of death for mice treated with ACA is equal from the first entry time to the last exit. We checked this assumption using a test of the correlation between the scaled Schoenfeld residuals and Kaplan-Meier transformed survival times (implemented as the `cox.zph` function in the survival package for R, [4]) and found no evidence for deviations for any of the design parameters nor for the included SCFAs.

Table B.4: Tests of non-proportionality of survival effects

Term	Correlation	P
propionate	-0.127	0.155
butyrate	0.039	0.633
acetate	0.012	0.885
ACA	0.047	0.630
female	0.065	0.518
UM	-0.036	0.701
ACA:female	-0.090	0.401
ACA:UM	0.044	0.643
female:UM	-0.028	0.778
ACA:female:UM	0.037	0.726
GLOBAL	—	0.922

Similarly, visual inspection of residual plots did not provide any evidence of deviations from linearity assumptions inherent to the model (Figure B.1).

While regression coefficients can be directly interpreted as a proportional increase or decrease in hazard of death at all time points, in the general case, this does not equate to a proportional change in expected survival time. It can therefore be challenging to understand the magnitude of the survival effect on expected lifespan. To provide a more intuitive demonstration of the size of the effect, we compared predicted survival curves for ACA treated, male mice at UM, with different SCFA concentrations characteristic of two existing individuals, based on a hypothetical scenario in which mice were alive at 830 days of age (i.e. a conditional expected survival curve; see Figure B.2). These simulated results demonstrate the strength of the association between SCFAs and survival over observed differences in concentrations.

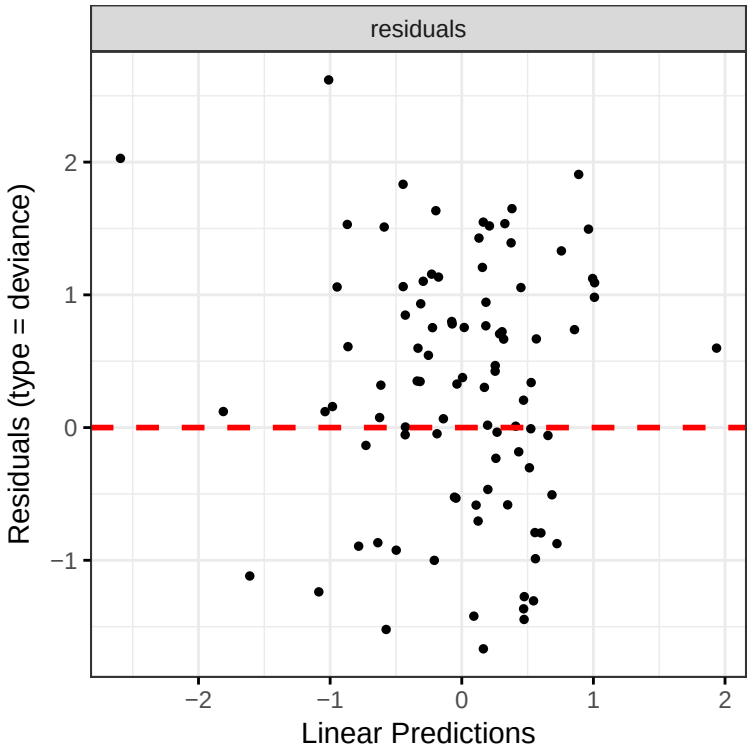

Figure B.1: Proportional hazard model residuals. Deviance residuals versus predicted log-hazard in a model of mouse survival that includes all design parameters (site, sex, and treatment) as well as the three SCFAs: propionate, butyrate, and acetate.

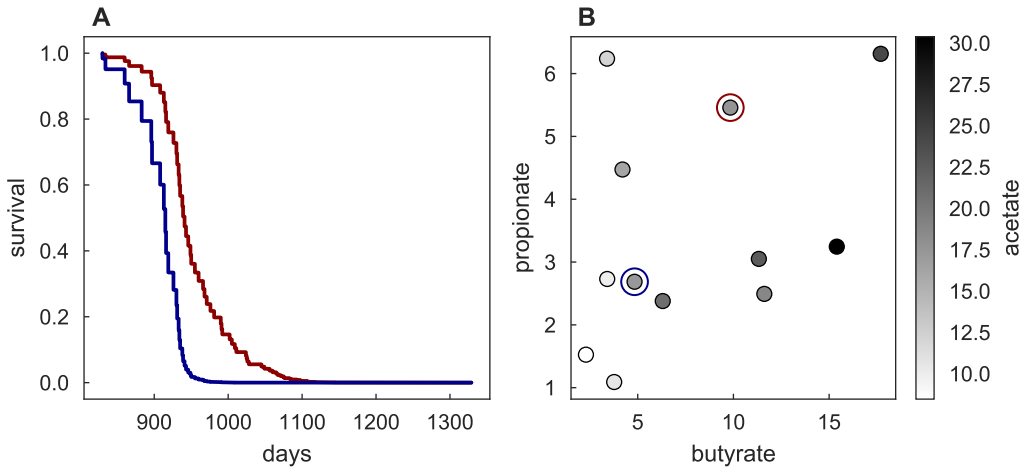

Figure B.2: Predicted effects of changes in SCFA concentration on mouse longevity. Survival of mice exhibiting realistic variation in SCFA concentrations. (**A**) Expected survival curves for male, ACA treated mice at UM, conditional on being alive at 830 days of age, and (**B**) SCFA concentrations for that same set of mice. Two representative butyrate, propionate, and acetate concentrations were chosen to match the measured concentrations for a high (red) and low (blue) butyrate/propionate individual, both having similar acetate concentrations.

BIBLIOGRAPHY

[1] Harrison, D.E., Strong, R., Allison, D.B., Ames, B.N., Astle, C.M., Atamna, H., Fernandez, E., Flurkey, K., Javors, M.A., Nadon, N.L., Nelson, J.F., Pletcher, S., Simpkins, J.W., Smith, D.L., Wilkinson, J.E., Miller, R.A.: Acarbose, 17-α-estradiol, and nordihydroguaiaretic acid extend mouse lifespan preferentially in males. Aging Cell **13**(2), 273–282 (2014). doi:10.1111/acel.12170

[2] Strong, R., Miller, R.A., Antebi, A., Astle, C.M., Bogue, M., Denzel, M.S., Fernandez, E., Flurkey, K., Hamilton, K.L., Lamming, D.W., Javors, M.A., de Magalhães, J.P., Martinez, P.A., McCord, J.M., Miller, B.F., Müller, M., Nelson, J.F., Ndukum, J., Rainger, G.E., Richardson, A., Sabatini, D.M., Salmon, A.B., Simpkins, J.W., Steegenga, W.T., Nadon, N.L., Harrison, D.E.: Longer lifespan in male mice treated with a weakly estrogenic agonist, an antioxidant, an α-glucosidase inhibitor or a Nrf2-inducer. Aging Cell **15**(5), 872–884 (2016). doi:10.1111/acel.12496

[3] Grambsch, P.M., Therneau, T.M.: Proportional hazards tests and diagnostics based on weighted residuals. Biometrika **81**(3), 515–526 (1994). doi:10.1093/biomet/81.3.515

[4] Therneau, T.M.: A Package for Survival Analysis in S (2015). https://cran.r-project.org/package=survival